Alfredo Cuzzocrea Christian Kittl
Dimitris E. Simos Edgar Weippl
Lida Xu (Eds.)

Availability, Reliability, and Security in Information Systems and HCI

LNCS 8127

IFIP WG 8.4, 8.9, TC 5 International
Cross-Domain Conference, CD-ARES 2013
Regensburg, Germany, September 2013, Proceedings

ifip

Springer

Lecture Notes in Computer Science 8127

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Alfredo Cuzzocrea Christian Kittl
Dimitris E. Simos Edgar Weippl Lida Xu (Eds.)

Availability, Reliability, and Security in Information Systems and HCI

IFIP WG 8.4, 8.9, TC 5 International
Cross-Domain Conference, CD-ARES 2013
Regensburg, Germany, September 2-6, 2013
Proceedings

Springer

Volume Editors

Alfredo Cuzzocrea
ICAR-CNR
and University of Calabria
Rende Cosenza, Italy
E-mail: cuzzocrea@si.deis.unical.it

Christian Kittl
Evolaris Next Level
Graz, Austria
E-mail: christian.kittl@evolaris.net

Dimitris E. Simos
SBA Research
Vienna, Austria
E-mail: dsimos@sba-research.org

Edgar Weippl
Vienna University of Technology
and SBA Research
Vienna, Austria
E-mail: edgar.weippl@tuwien.ac.at

Lida Xu
Old Dominion University
Norfolk, VA, USA
E-mail: lxu@odu.edu

ISSN 0302-9743
ISBN 978-3-642-40510-5
DOI 10.1007/978-3-642-40511-2
Springer Heidelberg New York Dordrecht London

e-ISSN 1611-3349
e-ISBN 978-3-642-40511-2

Library of Congress Control Number: 2013945883

CR Subject Classification (1998): H.4, K.6.5, H.3.1, H.3.3-5, E.3, K.4.4, H.2.3, H.2.8, H.5.3, D.2, J.1

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The Cross-Domain Conference and Workshop CD-ARES is focused on the holistic and scientific view of applications in the domain of information systems.

The idea of organizing cross-domain scientific events originated from a concept presented by the IFIP President Leon Strous at the IFIP 2010 World Computer Congress in Brisbane, which was seconded by many IFIP delegates in further discussions. Therefore, CD-ARES concentrates on the many aspects of information systems in bridging the gap between the research results in computer science and the many application fields.

This effort leads us to the consideration of the various important issues of massive information sharing and data integration, which will (in our opinion) dominate scientific work and discussions in the area of information systems in the second decade of this century.

The organizers of this event who are engaged within IFIP in the area of Enterprise Information Systems (WG 8.9), Business Information Systems (WG 8.4), and Information Technology Applications (TC 5) very much welcome the typical cross-domain aspect of this event.

The collocation with the SeCIHD 2013 Workshop was another possibility to discuss the most essential application factors. Special thanks to Professor Ilsun You for all his efforts in this special track, which was this year for the third time.

Also, we are proud to announce the Special Session Human–Computer Interaction & Knowledge Discovery (HCI-KDD), which was organized in the context of CD-ARES 2013. The ultimate goal of the task force HCI-KDD is to combine the best of two worlds: human–computer interaction (HCI), with emphasis on human intelligence, and knowledge discovery from data (KDD), dealing with computational intelligence. The cross-domain integration and appraisal of different fields provide an atmosphere in which to foster different perspectives and opinions. Special thanks to Dr. Andreas Holzinger, who made it possible to bring together researchers from diverse areas in a highly inter-disciplinary manner, to stimulate fresh ideas and encourage multi-disciplinary work.

Today, e-business depends heavily on the major cryptographic breakthroughs of almost 40 years ago. Without asymmetric cryptography, hardly any form of business transaction would be as easy to secure as it is today. We are thus very happy to have an excellent section on applied cryptography in this book.

The special track on modern cryptography and security engineering (MoCrySEn) attracted 30 submissions, of which the Program Committee selected 16 for publication in the workshop proceedings. The accepted papers dealt with symmetric-key cryptography, public-key cryptography, algorithmic cryptanalysis, software and hardware implementation of cryptographic algorithms, database encryption and interaction between cryptographic theory and implementation issues.

The papers presented at this conference were selected after extensive reviews by the Program Committee with the essential help of associated reviewers.

We would like to thank all the Program Committee members and the reviewers, who made great effort contributing their time, knowledge, and expertise and foremost the authors for their contributions.

September 2013

Alfredo Cuzzocrea
Christian Kittl
Dimitris E. Simos
Edgar Weippl
Lida Xu

Organization

Cross-Domain Conference and Workshop on Multidisciplinary Research and Practice for Information Systems (CD-ARES 2013)

General Chairs

Alfredo Cuzzocrea	University of Calabria, Italy
Edgar Weippl (IFIP WG 8.4 Chair)	Vienna University of Technology, Austria

Program Committee Chairs

Lida Xu (IFIP WG 8.9 Chair)	Old Dominion University, USA
Christian Kittl (IFIP WG 8.4)	Evolaris next level, Austria

Program Committee

Cristina Alcaraz Tello	University of Malaga, Spain
Markus Aleksy	ABB AG, Germany
Amin Anjomshoaa	Vienna University of Technology, Austria
Esma Aimeur	University of Montreal, Canada
Marianne Azer	National Telecommunication Institute and Nile University, Egypt
Kristian Beckers	University of Duisburg-Essen, Germany
Svetla Boytcheva	Bulgaria
Elzbieta Bukowska	Uniwersytet Ekonomiczny w Poznaniu, Poland
Lois Burgess	University of Wollongong, Australia
Stephan Faßbender	University of Duisburg-Essen, Germany
Meliha Handzic	Suleyman Sah University, Turkey
Stefan Hofbauer	Amadeus Data Processing GmbH, Germany
Andreas Holzinger	Medical University of Graz, Austria
Stefan Jakoubi	Security Research, Austria
Shinsaku Kiyomoto	KDDI R&D Laboratories Inc., Japan
John Krogstie	Norwegian University of Science and Technology, Norway
Javier Lopez Munoz	University of Malaga, Spain

Massimo Mecella	Sapienza Università di Roma, Italy
Bela Mutschler	University of Applied Sciences Ravensburg-Weingarten, Germany
Grant Osborne	Australian Goverment, Australia
Richard Overill	King's College London, UK
Andreas Peter	University of Twente, The Netherlands
Nalinpat Porrawatpreyakorn	King Mongkut's University of Technology North Bangkok, Thailand
Reinhard Riedl	Bern University of Applied Sciences, Switzerland
Simon Tjoa	St. Poelten University of Applied Sciences, Austria
A. Min Tjoa	Vienna University of Technology, Austria
Tetsuya Uchiki	Saitama University, Japan
Kangbin Yim	SCH University, Korea

Special Session on Human–Computer Interaction and Knowledge Discovery (HCI-KDD 2013)

Chairs

Andreas Holzinger	Medical University of Graz, Austria
Matthias Dehmer	UMIT Tyrol, Austria
Igor Jurisica	IBM Discovery Lab, Canada

International Scientific Committee

Amin Anjomshoaa	Vienna University of Technology, Austria
Matthieu d'Aquin	The Open University, Milton Keynes, UK
Joel P. Arrais	University of Coimbra, Portugal
Mounir Ben Ayed	Ecole Nationale d'Ingenieurs de Sfax, Tunisia
Matt-Mouley Bouamrane	University of Glasgow, UK
Polo Chau	Carnegie Mellon University, USA
Chaomei Chen	Drexel University, USA
Nilesh V. Chawla	University of Notre Dame, USA
Tomasz Donarowicz	Wroclaw University of Technology, Poland
Achim Ebert	Technical University of Kaiserslautern, Germany
Max J. Egenhofer	University of Maine, USA
Kapetanios Epaminondas	University of Westminster, London, UK
Massimo Ferri	University of Bologna, Italy
Alexandru Floares	Oncological Institute Cluj-Napoca, Romania
Ana Fred	IST – Technical University of Lisbon, Portugal
Adinda Freudenthal	Technical University Delft, The Netherlands
Hugo Gamboa	Universidade Nova de Lisboa, Portugal
Venu Govindaraju	University of Buffalo State New York, USA
Gary Gorman	Asia-New Zealand Informatics Associates, Malaysia
Michael Granitzer	University Passau, Germany
Dimitrios Gunopulos	University of Athens, Greece
Siegfried Handschuh	Digital Enterprise Research Institute, Ireland
Helwig Hauser	University of Bergen, Norway
Jun Luke Huan	University of Kansas, Lawrence, USA
Anthony Hunter	UCL University College London, UK
Alfred Inselberg	Tel Aviv University, Israel
Kalervo Jaervelin	University of Tampere, Finland
Igor Jurisica	IBM Life Sciences Discovery Centre & University of Toronto, Canada

Jiri- Klema	Czech Technical University, Prague, Czech Republic
Lubos Klucar	Slovak Academy of Sciences, Bratislava, Slovakia
David Koslicki	Pennsylvania State University, USA
Patti Kostkova	City University London, UK
Damjan Krstajic	Research Centre for Cheminformatics, Belgrade, Serbia
Natsuhiko Kumasaka	Center for Genomic Medicine (CGM), Tokyo, Japan
Nada Lavrac	Joszef Stefan Institute, Ljubljana, Slovenia
Pei Ling Lai	Southern Taiwan University, Taiwan
Alexander Lex	Harvard University, Cambridge (MA), USA
Chunping Li	Tsinghua University, Canada
Luca Longo	Trinity College Dublin, Ireland
Lenka Lhotska	Czech Technical University Prague, Czech Republic
Andras Lukacs	Hungarian Academy of Sciences and Eoetvos University, Budapest, Hungary
Avi Ma' Ayan	The Mount Sinai Medical Center, New York, USA
Ljiljana Majnaric-Trtica	Josip Juraj Strossmayer University, Osijek, Croatia
Martin Middendorf	University of Leipzig, Germany
Silvia Miksch	Vienna University of Technology, Austria
Antonio Moreno-Ribas	Universitat Rovira i Virgili, Tarragona, Spain
Marian Mrozek	Jagiellonian University, Krakow, Poland
Daniel E. O'Leary	University of Southern California, USA
Ant Ozok	UMBC, Baltimore, USA
Vasile Palade	University of Oxford, UK
Jan Paralic	Technical University of Kosice, Slovakia
Valerio Pascucci	University of Utah, USA
Gabriella Pasi	Università di Milano Bicocca, Italy
Margit Pohl	Vienna University of Technology, Vienna, Austria
Paul Rabadan	Columbia University College of Physicians, New York, USA
Heri Ramampiaro	Norwegian University of Science and Technology, Norway
Dietrich Rebholz	European Bioinformatics Institute, Cambridge, UK
Gerhard Rigoll	Munich University of Technology, Germany
Lior Rokach	Ben-Gurion University of the Negev, Israel
Carsten Roecker	RWTH Aachen University, Germany
Giuseppe Santucci	La Sapienza, University of Rome, Italy

Reinhold Scherer	Graz BCI Lab, Graz University of Technology, Austria
Paola Sebastiani	Boston University, USA
Christin Seifert	University of Passau, Germany
Tanja Schultz	Karlsruhe Institute of Technology, Germany
Andrzej Skowron	University of Warsaw, Poland
Rainer Spang	University of Regensburg, Germany
Neil R. Smalheiser	University of Illinois at Chicago, USA
Marc Streit	Johannes-Kepler University Linz, Austria
Dimitar Trajanov	Cyril and Methodius University, Skopje, Macedonia
A Min Tjoa	Vienna University of Technology, Austria
Olof Torgersson	Chalmers University of Technology and University of Gothenburg, Sweden
Patricia Ordonez-Rozo	University of Maryland, Baltimore County, Baltimore, USA
Jianhua Ruan	University of Texas at San Antonio, USA
Pak Chung Wong	Pacific Northwest Laboratory, Washington, USA
William Wong	Middlesex University London, UK
Kai Xu	Middlesex University London, UK
Pinar Yildirim	Okan University, Istanbul, Turkey
Martina Ziefle	RWTH Aachen University, Germany
Ning Zhong	Maebashi Institute of Technology, Japan
Xuezhong Zhou	Beijing Jiaotong University, China

International Industrial Applications and Business Committee

Peter Bak	IBM Haifa Research Lab, Mount Carmel, Israel
Andreas Bender	Unilever Centre for Molecular Science Informatics, Cambridge, UK
Alberto Brabenetz	IBM Vienna Austria, Austria
Anni R. Coden	IBM T.J. Watson Research Center New York, USA
Stefan Jaschke	IBM Vienna Austria, Austria
Homa Javahery	IBM Centers for Solution Innovation, Canada
Igor Jurisica	IBM Life Sciences Discovery Center, Canada
Mei Kobayashi	IBM Tokyo Laboratory, Tokyo, Japan
Alek Kolcz	Twitter Inc., USA
Jie Lu	IBM Thomas J. Watson Research Center, Hawthorne, New York, USA
Helmut Ludwar	IBM Vienna, Austria
Sriganesh Madhvanath	Hewlett-Packard Laboratories, Bangalore, India
Roberto Mirizzi	Hewlett-Packard Laboratories, Palo Alto, USA

Laxmi Parida	IBM Thomas J. Watson Research Center, New York, USA
Hugo Silva	PLUX Wireless Biosensors, Lisbon, Portugal
Gurjeet Singh	Ayasdi, USA
Dan T. Tecuci	Siemens Corporate Research, Princeton, USA
Uli Waltinger	Siemens Business Analytics, Germany
Raffael Wiemker	Philips Research Hambug, Gernany
Minlu Zhang	Nextbio, USA

International Student's Committee

Andre Calero-Valdez	RWTH Aachen, Germany
Pavel Dlotko	Jagiellonian University, Poland
Markus Fassold	hic4all Team Graz, Austria
Fleur Jeanquartier	hci4all Team Graz, Austria
Emanuele Panzeri	University of Milano-Bicocca, Italy
Igor Pernek	University of Maribor, Slovenia
Vito Claudio Ostuni	Politecnico di Bari, Italy
Christof Stocker	hci4all Team Graz, Austria
Hubert Wagner	Jagiellonian University, Poland

Table of Contents

Cross-Domain Conference and Workshop on Multidisciplinary Research and Practice for Information Systems (CD-ARES 2013)

Economic, Ethical, Legal, Multilingual, Organizational and Social Aspects

Context-Oriented Information Integration

Data / Information Management as a Service

Context-Oriented Information Integration and Location-Aware Computing

Security and Privacy

Risk Management and Business Continuity

Security and Privacy and Location Based Applications

Human–Computer Interaction and Knowledge Discovery (HCI-KDD 2013)

Sensor Data Meets Social Networks Reflecting on Benefits in the Case of a Patient Room

Fabienne Kuhn, Andreas Spichiger, and Reinhard Riedl

Bern University of Applied Sciences, Bern, Switzerland
{fabienne.kuhn,andreas.spichiger,reinhard.riedl}@bfh.ch

Abstract. In a hospital, information exchange is essential to save lives and to prevent life-endangering mistakes. Information exchange is supported by a hospital information system (HIS). From a theoretical perspective, the deployment of an HIS is promising because it reduces errors and duplication of information. In practice, however, there are some major problems concerning the usage of such a system. One way to deal with these problems is introduced in this paper: the integration of sensor data into social media. The paper concentrates on the conceptual benefits and risks such an integration may generate. It focuses on the case of a patient room.

Keywords: Social media, sensor media, requirements engineering, ubiquitous computing, information provider, information access, Web 2.0, E-health.

1 Introduction

The information management in a hospital is often supported by a hospital information system (HIS) which addresses the various needs in a hospital by integrating most applications used there [10]. HIS provides some essential benefits, but there can still be major problems when using it on a daily basis. Benefits of using an effective HIS include information integrity, prevention of medication errors, reduction of transcription errors, duplication of information entries and report turnaround time [10], [12]. Problems arise in the following areas [33], [39], [30], among others:

Data recording and storage

- Immediate data access is required in a context where urgent cases are commonplace. Depending on the data storage solution, this requirement may not be met because data cannot always be recorded and stored in a timely manner. Sometimes not only the data storage but also the workflow does not allow instant recording of data. Concerns arise that information does not flow quickly enough during personnel changeovers, e.g. a change of shift.
- When a patient moves from one ward to another there are no techniques to move his data along effectively. This is because one organization unit cannot copy just one patient's data from another organization unit. Furthermore it is not easy to find old data about a newly arrived patient, for example when this person comes from an external domain.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 1–15, 2013.

- Classic information systems are constructed for people who work on computers and nursing staff work with humans. The time taken to record data sometimes is longer than is allowed for in a nurse's job description. This leads to poor recording habits like copy/paste from other reports or shortening notes so they become unspecific. Thus wrong medication[1] and wrong therapy can result.

Data view

- Doctors and nursing staff can lose the overview because there are many different screens where many details can be entered, which are not necessarily important but can be shown anyway.
- When data is not filtered and not declared, doctors and nursing staff may focus on specifics and do not see what others have written.
- Some approaches use alerts, reminders and warning messages to try to generate a specific view for each user. However, they can be sent regardless whether the information is relevant for the recipient or not. Thus the users tend to turn them off.

Culture

- Communication, which is very important in a hospital, is not simply about exchanging information. It is about generating effects; the sender wants the recipient to act. Furthermore, with direct communication the sender can find out how much the recipient understands, can correct a misunderstanding and can estimate and influence the recipient's actions. It is also about maintaining relationships.

We are interested in how user interfaces can be rebuilt as a smart system, focusing mainly on two problem areas. An important area concerning the problems mentioned earlier is the recording of data. It should be correct but not time consuming; it should support the staff but not interfere significantly with their workflow. A possible solution can be the deployment of sensors. Nevertheless it is risky to add more information into a system which already has several unresolved issues. The expectation is that it will not work. Let us follow this idea anyway: how should a system be rearranged to improve its usage? For a start there should be sensors which record the well-being of a patient and his whereabouts. Furthermore, medicines should be locatable to determine which medicine was given to whom. Another important problem area is the data view. The view should be composed user-and-situation-specific way without irrelevant data. Social media can provide those views. It also addresses another problem with HIS: data privacy. Social media offers the possibility to share selected data only with a selected group of people. Along with the previously mentioned requirements, which can be met either by sensor or social media, others can be met by the combination of both. For example the actors may comprehend the context and orient themselves better. This paper shows further possible conceptual benefits and risks concerning the idea of the combination of sensor and social media (SoSeMe) as a possible solution to the problems mentioned above. We have concentrated on the case of a patient room. This case was selected because a patient room provides a small and controlled environment and different stakeholders meet multifaceted challenges.

[1] See also [41].

Fig. 1 gives an overview of the subjects raised thereby. Our research follows research guidelines according to Hevner [32]. In compiling the results we have mainly focused on the dimensions quality and economy. The paper is structured as follows: Theoretical basics and ideas are provided in chapter 2. Chapter 3 summarizes the case results and shows topics discussed for the patient-room case.

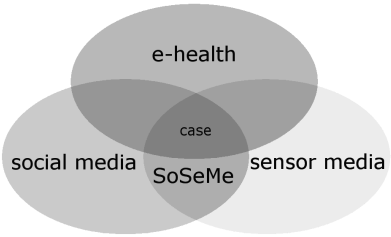


Fig. 1. Situation of the subject of this paper

Related Research Work
The idea of combining social and sensor media is not new. When users are requested to tag their posts with keywords, the benefits increase because it is easier to find posts related to specific topics [15]. An example here is Twitter, there the user can tag keywords by typing a "#" at the beginning. This makes it easier for machines to find relevant related information for analysis. For example natural disasters are easier to detect and to monitor by mining the posts of social media users when they distinguish keywords like the name of the service which analyzes the data [16]. Areas where social and sensor networks can be linked effectively are shown in [36]. The article concludes that SoSeMe applications are more context aware. They are more deeply integrated into daily life and can act as an enabler for emerging services, e.g. mobile communication, internet, multimedia, etc. The framework in [37] describes communication in e-health primarily through social media. Those mentioned papers clearly indicate potential benefits. Even so, SoSeMe contrasts itself by researching in an area where data is mainly gathered through sensors, not people, and stored within the logic of social media. Furthermore the communication processes and rituals in a patient room are examined. With this approach, advantages in the limited setting and the intricacies of traditional communication are well understood.

2 Data Processing

Sensor and social media have different roles to play. In the following chapters their relevant roles are introduced for the case. This short theoretical background focuses on data processing: how data is acquired, stored and viewed. Furthermore it introduces topics inevitably connected with either social or sensor media in the context of a patient room.

2.1 E-Health

E-health focuses on the technical side of electronic media and adapts it to health care [17]. Health 2.0 uses networks of stakeholders like patients, doctors and scientists to

personalize health care and promote collaboration [18]. It allows the patients to actively participate in their own healthcare through information and communication technology. In Fig. 1 Health 2.0 is the interface between e-health and social media.

Health networks may have the same problem as large enterprises: they are extremely complex and different parts often operate in silos [19]. This makes it hard to communicate across organizational units. However, if the flow of information between them stagnates, significant risks would be created for the patients. In addition, the non-sharing of information increases the risk of reacting too late to the spreading of infectious diseases. One solution to the latter problem is found in the Global Pandemic Initiative from 2006, where IBM offers several pieces of its healthcare technology portfolio to work with major worldwide public institutions [20]. Technologies offered include a framework to improve communication and collaboration, building a community of users which can tap into information as well as conducting research on influenza viruses. A solution for the former disadvantage may be found in patient centered electronic health records (EHR). EHR offers an overview of important medical data about a patient. The patient decides himself who gets access to his data.

2.2 Sensor Media

Sensors record their surroundings on demand, regularly or permanently within a time period. They can determine the position, movement and further attributes of an object at a specific time. They can also detect pressure to find out the weight of an item or detect the movements of a person during the day, categorizing them into different activities [34]. Sensors can be put together into a network in which they communicate with each other [8], [25]. The combination of networked sensors opens new possibilities and more attributes can be measured. Correlation allows a more precise acquisition of an object's state and it is possible to see the bigger context. In the patient room the combination of sensors in a network can, for example, detect that the patient fell in the wet room at 12:00 p.m. Sensor media record an event but with sensors alone there is no reaction. How does the nurse know a patient has had a fall?

2.3 Sensor Deployment in Healthcare

The deployment of sensors in healthcare has five main application areas [21]: process controlling and documentation, localization, personalized patient medicine and identification, monitoring of measured data and protection against fake medicine.

In *process controlling*, the sensor technology can facilitate processes and organize them more efficiently based on data collected through sensors. This can optimize product logistics and cost control. Processes concerning documentation can also be simplified, which results in timesaving and reduction of mistakes. Through a better *localization* of personnel, materials and devices, search time can be reduced and information about the inventory can be provided. One of the most frequent mistakes is the dispensing of the wrong medicine. Insufficient information about the exact kind, amount and time the medicine was administered is one of the causes. A personalized RFID tag can prevent such mistakes, because with this, all information can be allocated to a specific patient. So it can be ensured that the right patient gets the right

medicine. The *measurement* and *monitoring* of vital signs is relevant for product logistics. The collected data is transferred into an emergency system that raises an alarm when a specific range is exceeded. *Protection* against counterfeit medicine is an issue whose relevance is increasing constantly. RFID tags enable the definite identification of medicines and permanent, gapless data alignment. This simplifies the distinction between fakes and originals.

2.4 Social Media

With social media people can share data, view data and use others' data. Users have the possibility of sharing data publicly or with a predefined specific group of people [11]. Shared data have different levels of quality. They can even be faked to generate a specific impression, for example to compete or to impress. Most social media further provide individual views of specific situations[2]. This means the information about a situation is selected to create specific, person-centered perspectives.

Social media can also be seen as a recorder for time-dependent distributed states [5]. The concept of the time line increases the quality of the data on social media because each data item gets an exact timestamp and orders data chronologically.

Social media are also able to acquire data about the relationship between persons, or persons and items, or even items and items, based on specific data which are uploaded by the people themselves [13]. The possibility to link data, e.g. by tagging photos with the persons present on them, has proven to be an important tool for relationship management. One of the more recent trends is that people tend to share and comment on data which they find important, rather than uploading data themselves [31].

With social media, the patient could inform the nursing staff that he fell down in the wet room at 12:00 p.m. But how does he do that when he is unconscious?

2.5 Social and Sensor Media Fusion

In SoSeMe, data which is collected and joined by sensor media are stored in and distributed by social media. Data is provided by the involved actors/stakeholders who write statements into social media, by the sensor network, and through the hospital information system. The social media set-up plus the attribution of metadata help to contextualize the information, customized for those who access it. This makes the information pick-up much easier and opens a whole new experience in data handling. In particular, both missing the pick-up of relevant information and giving wrong interpretation to the accessed data become less likely. While the set-up particularly addresses humans, it could equally be employed to steer actuators based on artificial intelligence.

Provided that the patient shared his data with nursing staff and that the patient room is equipped with sensors, the data collected when he fell down in the wet room at 12:00 p.m. can be stored in social media and the nursing staff will be informed and can provide immediate help.

---

[2] For further information about an instrument how this could work see `https://developers.facebook.com/docs/technical-guides/fql/`

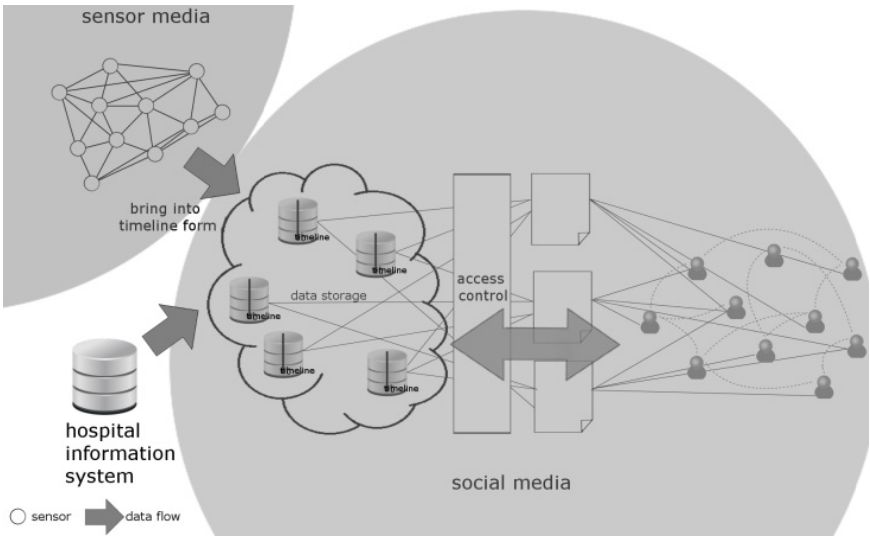


Fig. 2. Interplay between social and sensor media

3 Case Patient Room

3.1 General Scenario

The activities in the patient room include people with mixed attributes like patients, visitors and health professionals who communicate with one another and interact with things. There are lots of situations when communication is of critical importance for the health of the patient. The intricacies and problems of this communication are well understood. On the one hand, the need for high quality integration of information is obvious; on the other hand a lot of this information concerns the privacy of patients, which has to be protected. A lot of different kinds of sensors may be placed to observe the room with all its content almost fully, but there is also a lot of social context information being created outside, which also plays a key role.

For a better understanding of the current situation, we have conducted qualitative, semi-structured interviews with the staff in hospitals. Furthermore the processes in a hospital were observed. Those empirical assessments were made according to Pohl [22] in the hospitals Spital Zentrum Biel, Spital Aarberg and Inselspital Bern. Those hospitals were selected to have a comparison between a small hospital which is typically equipped, one that has recent medical equipment as well as infrastructure and a large university hospital. The empirical assessments where focused on the available equipment, available information, nursing and information processes. To evaluate the results, the software MaxQDA 10[3] was used. Parallel to and based on the empirical assessments we have determined three specific indicative scenarios for investigation (cf. chapter 3.2), for which we then have identified the optative scenarios and the

[3] For further information about this software see `http://www.maxqda.de`

related requirements also according to Pohl. This has yielded ideas about how to conceptually optimize patient rooms in the future, which we explain in chapter 3.3. Then we have built a prototype of a future patient room equipped with a sensor network.

Different prototype implementations were used to find out what is technically possible. Thus another scenario was evaluated. Active sensors were used to identify doctors and nursing staff, patients and medicine. The nurse carried a small device which showed information relevant to the room s/he was in. When the nurse had the wrong medicine for a patient, the device would give a warning. The drawback here is that the prototype reacted really slowly, meaning that the sensor network has to be optimized. This paper does not go deeper into the results generated by the prototype, because they only provide a basis for further project ideas.

3.2 Patient Room Today

We looked at the situation of information management in three hospitals in and around Bern, Switzerland. Each had differently equipped patient rooms. However, in all cases there was a paging system, an intravenous pole, an intravenous bag, an infusion set, an Infusomat and a bed.

With the *patient paging system* the patient can call for a nurse, which simplifies communication between the patient room and the nurses' room. This system consists of a button or a string near the bed, the shower, the toilet and the eating table. When a patient needs help, he pushes the button. Then a lamp turns on beside the door. In the corridor and nurses' room is a display which shows in which room the button was pushed. After entering the room, the nurse turns off the lamp beside the door, so other nurses know that this patient is already being attended to. The nurse can also signal an emergency through this system. Then the door lamp shows the letter N (for 'Notfall', which is German for emergency).

Intravenous bags contain the infusion solution. The name of the infusion and additional information about the content and context (e.g. the nurse's signature) of the bag are written on the bag.

Infusomats[4] are electronic pumps which ensure the supply of liquid drugs during longer periods. Each Infusomat has a controlled hose clip with which an exact dosage for the infusion solution per interval is possible. They show the total amount of infusion solution which has already passed through, how many milliliters per hour should be administered and how many milliliters have already been administered. Some Infusomats have an integrated sensor which counts each drop. When all of the infusion solution has passed through, the device reports the end of the infusion.

Infusions can cause problems like vein inflammation which cannot be detected immediately. Furthermore it is sometimes not possible to determine whether the intravenous catheter is set right into the vein, which could inhibit blood flow.

In several scenarios the bed plays a central part; from here the patient interacts with his environment. As a minimum, the height and the angle of the bedhead are adjustable. In the *bed* the patient can be transported. In most cases the bed can be extended.

[4] Infusomat is a trademark of B. Braun Melsungen AG.

In a patient room there are three main groups of persons: patient, staff and visitors. The patient suffers from illness or the consequences of an accident. He is identified through his name and forename which are written on the bed and the wardrobes. In the researched context, the staff mainly consists of the doctors and nursing staff, which provide services to the patients. They wear a badge on which relevant metadata about their person is written. The doctors prescribe drugs and enter the prescriptions into the hospital information system.

Finally, time management is an important issue. For example there is a schedule for the visits of the chief physician. The use of digital information processing systems is still rather limited. A lot of the information processed only exists in the heads of the staff and is communicated mostly in face-to-face meetings.

3.3 Patient Room in the Future

In the following paragraphs the vision of the interviewed persons about the future patient room is summarized and adapted to the main aspects for this case.

The patient room of the future is a smart system and in this case is called *smart patient room system* (SPRS). It monitors and controls the patient's situation because it has the best access to all relevant information. SPRS is the specific product of SoSeMe for the case patient room.

The *patient page system* is automated through new technologies. The HIS detects critical situations on the basis of sensor data and sends for a nurse who herself has an individual device which shows her relevant patient data.

The *intravenous pole* recognizes infusion and drugs. The SPRS compares these with the patient data. If there is a prescription for the correctly identified patient, it starts the administration of the infusion. If any error occurs, the system will warn the nurse about the danger.

The catheter of an *infusion* could be improved by adding a sensor which can determine if it is properly in the vein or not.

The *bed* is intelligent, but still easy to use and comfortable. It records the patient's vital signs, body temperature, blood pressure and weight repeatedly and sends the data through a communication channel to social media storage.

Consumables like infusion bags and medicine are equipped with RFID-tags.

3.4 Specific Scenario Versions

Throughout the interviews, with nurses working at different hospitals and an M.Sc. class of nurses, we assessed the situation in standard patient rooms and the problems arising because of lack of easily accessible information. Out of the large number of possible scenarios three were selected. The specific scenario '**Infusomat**' was selected because it addresses major issues (identifying the patient, dispensing of the wrong medicines) and also includes an actuator, as well as the necessary sensors. The specific scenario '**Vital Signs**' was selected because this scenario happens frequently and integration of information flow would improve the work significantly. The specific scenario '**Visitors**' was selected to show how people get involved in information

diffusion in a patient room. The goal of the selection was to have a minimal set of scenarios that allow all the important benefits and risks to be demonstrated. After the description of today's process of the indicative scenarios, they were optimized as optative scenarios to fit in the future patient room, as shown below. The prototype of the future patient room was then built for optimal support of these scenarios.

Infusomat
A patient is suffering from a chronic lung disease and needs an analgesic. His nurse orders an infusion over the SPRS by voice command through her headset. She arranges the printing of an RFID tag, which she uses to get the analgesic and infusion from the medicine cabinet. She sticks the tag on the infusion bag and adds the analgesic to the infusion. The infusion bag is attached and integrated to the smart infusion booth (SIB), an intravenous pole with built-in electronics. The SPRS reads both SIB and RFID tag, identifies the patient and checks for a match between the data and the prescription. The SPRS gives the order to SIB to adjust. The SIB gives the signal for the infusion. The nurse attaches the infusion to the patient and starts the SIB. After an hour, when the infusion is finished, the SIB sends a signal to the SPRS that the infusion is finished. The SPRS informs the nurse with a voice mail. Then it turns itself off. The nurse removes the infusion.

Vital Signs
During the first four hours after an operation, the patient's vital signs, like blood pressure, temperature, oxygen saturation and wakefulness are determined and added to his electronic health records (EHR) every half an hour. In addition, the nurse asks the patient about nausea, vomiting, pain and the urge to void. The nurse appends the collected data to the EHR with the aid of a mobile device. The data can be accessed by doctors via social media, allowing better treatment if the patient has agreed to share EHR with them.

Visitors
There are two patients in the room. One has company from his wife, the other from a coworker. The SPRS identifies the wife and the coworker by video records. If the visitors wish it, SPRS creates a digital relationship between those two and connects this with various additional data, such as the nature of the relationship between patient and visitor (e.g. the relation between relatives and patients may be stronger than the one between coworkers and patients).

3.5 Potential Benefits

The analysis of the three specific versions of the more general patient room scenario have shown different areas where the use of sensor data and the providing of the necessary information in the right form by social media can produce benefits, mainly in the areas of quality and economy. A summary including stakeholders' perspectives is shown below. Note that the stakeholder list is not complete. Other stakeholders outside the patient room which can benefit from SoSeMe include hospitals, the pharmaceutical industry, health insurers, researchers and governments.

Patient and Medicine Identification Improves Medicine Management

- The system has the possibility to align the medicine and patient data.

Nursing Staff	• Nurses are prevented from making mistakes, because the system warns if someone tries to give the wrong medication. • They can provide the right amount of medicine in a timely manner.
Patients	• The patients get the correct treatment.

Automated, Real-time Information Exchange around the Clock Improves Communication and Data Flow

- Information and experiences with specific situations can be exchanged and shared free of barriers and media disruption.
- Communication is quick, easy and location- as well as time-independent.
- The system supports the nursing stuff to explain the patients' health in an understandable but correct way.
- Stakeholder interactions are improved by the exchange of opinions based on data.
- There is the possibility of location- and time-independent patient monitoring.

Nursing Staff	• Immediate information access in case of an emergency. • The insight of the patient's relatives can systematically be collected. • They can improve checklists about symptoms, which are relevant for the classification of appropriate therapy.
Patients	• The patients can participate in the decisions doctors render. • They get more secure treatments. • They can move more freely.
Relatives	• They can seek support from other relatives.

Automatic Information Acquisition Improves Stakeholder Management

- It is easier to perform analysis based on information about stakeholders. The system allows an easier collaboration.

Nursing Staff	• Staff can be deployed and scheduled more effectively. • Nurses can organize their nursing services better.
Patients, Relatives	• They get faster and more correct information.

Sensor Improve Workflow Facilitation

- Time-consuming and complex tasks can be facilitated.

Nursing Staff	• Data input through headsets instead of typing. • Nurses can organize their nursing services better.
Patients	• Record symptoms at moment of event.

3.6 Potential Risks

The cooperation of patients is needed for the success of the idea of SoSeMe. However, due to data protection concerns, patients may tend to try to hide things from their doctors [34] and manipulate sensors on purpose. This strategy may have consequences for the quality of the treatment. Moreover, doctors too may have the impression that their work would be checked more than before and they would have to justify themselves. Because of this feeling, doctors may tend to concentrate on the awareness of third parties rather than their patient's wellbeing.

With the social media in SoSeMe the patient is the data owner and can, as with the EHR, decide who gets access. The idea of SoSeMe depends on the patient's willingness to share his data and also his trust in technology. The amount of information available for the stakeholders increases with the integration of social and sensor media, meaning that analysis of this machine-sourced data could be invalid unless the significance of each piece of information is considered first. This argument is based on the assumption that people weight the information they communicate through speech by variation of pitch level and volume. Also the impression may arise that knowledge about the context is fully available and the importance of a piece of information is weighted differently by different stakeholders. Decisions can be made based on this impression. Furthermore, the nursing staff and doctors have to decide how much they want to know from their patients because of information processing costs and the risk increases with too much information.

Filters and access control may not work reliably or stakeholders abuse their information access. Therefore it is possible that people get access to data which they should not see. Furthermore, should patient data transfer not be encrypted as required by law [35], the risk of unauthorized data access and data abuse would increase. The usability of input devices is a further aspect which influences the quality of information in the system and the costs of information processing. As long as information has to be typed in rather than told to somebody, more time is needed or less information can be recorded. On the other hand, audio records cannot be processed optimally by machines. This introduces new areas of error.

3.7 Discussion

A future patient room should be as smart as possible so that the patient has more free space and can communicate with fewer barriers. The room should offer the nursing staff facilities within their workflow as well as location- and time-independent communication for patients, staff and visitors.

The combination of social and sensor media can also be used as an emergency system, because it can respond in a timely manner [23]. No matter how promising the integration of sensor data from a patient room into a social network is, it depends on the patient's maturity and sense of responsibility to himself. The interviews we conducted show that the integration of sensor technology can improve workflows. However, the interviewees also made clear that the resulting free capacity would be best used to increase the social interaction between nursing staff and patients and not cost cutting.

The main challenge faced by SoSeMe is data privacy. Interviewees from a technical department declared they would never publish health relevant data. This statement is comforting regarding privacy protection but many people already use online forums or social media to seek help for their illnesses [1]. We analyzed the relevant laws in Switzerland. The nursing staff is not allowed to distribute personal data without the permission of the patient[5]. They must not give access to third parties without the permission of the patient[6] and it is forbidden even to process the data unless the patient agrees to it first[7]. The patient is the data owner. This means the nursing staff has to be given access to this data and the patient decides who can access that data. With this, the main challenge faced by SoSeMe may be the main benefit, because SoSeMe permits the patients exactly what is stated by the law: an instrument with which they can act as data owner and easily give permission to organizations or persons to access data, process data or deny access altogether. Furthermore, through strict integration of social and sensor media data, privacy in systems can even be improved because violations of data privacy can be tracked.

SoSeMe includes the benefits of the patient's dossier (EHR[8]) and enriches it. In today's e-health concepts, it is envisioned that patients will take more responsibility for their own dossiers [26], with hospitals providing added value to their patients by supplying their personal health data in a suitable way. Each patient then has the possibility to share his EHR with whomever, whenever and wherever they want, which includes non-sharing as well. The patient may share data related to their current visit, or if necessary, up to all of their recorded history, as their care requires.

The patient's room situation in a SoSeMe setting is a very complex system. The requirements elicitation on the basis of three rather simple indicative scenarios, defining the indicative information model, and then elaborating the corresponding operative scenarios and the future information model was shown to be a very direct way to define the future system. The applied requirements elicitation techniques in combination with stakeholder interviews proved to be a very effective method. In combination they are appropriate for a discussion on the high abstraction level of potential directions of future systems.

4 Conclusion and Further Activities

Our research has shown that the integration of digital social media and of sensor networks offers a rich set of options for improving the communication taking place in and around a patient room. At the same time it has revealed significant risks if such an integration is not handled with care. In addition, it may be expected that such experiments would face significant opposition and ethical concerns. Nevertheless the combination of social and sensor media can provide a possible solution to problems arisen in the deployment of a HIS.

---

[5] Related articles are amongst others SR 101 Art. 13, SR 235.1 Art. 1 and 35

[6] Related article is SR 235.1 Art. 8

[7] Related article is SR 235.1 Art. 17

[8] For further information on a patient dossier in Switzerland see `http://www.evita.ch/en/home/`.

Potential benefits include the deployment of SoSeMe within the patient room, which may reduce wrong medication; patients can be monitored in real time and also be treated immediately in case of an emergency. Furthermore, medicine and personnel management can be improved, and workflow facilitated. Participation of patients in decision making can also be supported. Moreover, the idea of time line and collaboration can be a benefit for worldwide research, because with social media it is easier to connect locally independent information. Note that nurses' responsibilities are not being delegated to machines. Machines only act as support.

Since these benefits are so important, we have decided to continue our research and to investigate the specific sub-scenarios described in more depth. The patient room can further be equipped with more sensors such as video, audio and wireless sensor systems, which would allow, for example, location and identification of objects. In the patient room this could be used for checking the inventory [27], [19]. To find out what the patient ate a photo can be taken; to find out how much he ate, the food can be weighed. Zooming out of the patient room, in the future we shall look further at the hospital context as a whole or even into larger medical systems, such as the healthcare of a specific region.

Acknowledgements. The first thanks go to Emine Seker, Nicola Schwery and Michael Ferreira who provided important content for this project. Further thanks go to Ross Bennie, Alessia Neuroni and Thomas Jarchow, who reviewed this paper.

References

[1] Chun, S., MacKellar, B.: Social Health Data Integration using Semantic Web. In: SAC 2012, Riva del Garda, Italy, March 25-29, pp. 392–397. ACM, New York (2012), http://dx.doi.org/10.1145/2245276.2245351

[2] Bry, F., et al.: Digital Social Media, Dagstuhl Manifesto. University of Munich, Germany (2010), http://www.pms.ifi.lmu.de/publikationen/PMS-FB/PMS-FB-2010-7/PMS-FB-2010-7-dagstuhl-manifesto.pdf

[3] Vera, R., Ochoa, S.F., Aldunate, R.G.: EDIPS: An Easy to Deploy Indoor Positioning System to support loosely coupled mobile work. Personal and Ubiquitous Computing 15(4), 365–374 (2011), http://dx.doi.org/10.1007/s00779-010-0357-x

[4] Kamei, K., et al.: Tagging Strategies for Extracting Real-world Events with Networked Sensors. In: ICMI 2007, Nagoya, Japan, November 15, pp. 35–42. ACM, New York (2007), http://dx.doi.org/10.1145/1330588.1330594

[5] Sawyer, T.: Facebook Timeline. Rough Notes 155(66), 68–69 (2012), http://search.proquest.com/docview/1010360195?accountid=15920

[6] Mullender, S.: Distributed Systems. ACM Press, United States of America (1993)

[7] Bantel, M.: Messgeräte-Praxis, Funktion und Einsatz moderner Messgeräte. Hanser Verlag, München Wien (2004)

[8] Dargie, W., Poellabauer, C.: Fundamentals of wireless sensor networks: theory and practice. John Wiley and Sons, Hoboken (2010)

[9] Swedberg, C.: Michigan Researchers Develop RFID-based Sensors to Measure Physical Activity. RFID Journal (2010), http://www.rfidjournal.com/article/view/7884/

[10] EHR Scope: Hospital Information Systems (HIS) (2013), http://www.emrconsultant.com/education/hospital-information-systems

[11] Luo, W., Liu, J., Liu, J.: An Analysis of Security in Social Networks. In: DASC 2009, Chongqing, China, December 12-14, pp. 648–651 (2009), http://dx.doi.org/10.1109/DASC.2009.100

[12] Watson, L.: Woman dies after replacement nurse gives her wrong drug during labout dispute. Daily Mail, September 26 (2011), http://www.dailymail.co.uk/news/article-2041884/Woman-dies-replacement-nurse-gives-wrong-drug-labour-dispute.html#axzz2KcPIH9Of

[13] Yang, C.C., et al.: Identifying Implicit Relationships between Social Media Users to Support Social Commerce. In: ICEC 2012, Singapore, Singapore, August 07-08, pp. 41–47. ACM, New York (2012), http://dx.doi.org/10.1145/2346536.2346544

[14] Sharp, J.: A Look at Social Media in Health Care – Two Years Later. iHealthBeat, May 17 (2012), http://www.ihealthbeat.org/perspectives/2012/a-look-at-social-media-in-health-care-two-years-later.aspx

[15] Nagarajan, M., Sheth, A., Velmurugan, S.: Citizen sensor data mining, social media analytics and development centric web applications. In: WWW 2011, Hyderabad, India, March 28-April 1, pp. 289–290. ACM, New York (2011), http://dx.doi.org/10.1145/1963192.1963315

[16] Aulov, O., Halem, M.: Human Sensor Networks for Improved Modeling of Natural Disasters. IEEE 100(10), 2812–2823 (2012), http://dx.doi.org/10.1109/JPROC.2012.2195629

[17] Della Mea, V.: What is e-Health (2): The death of telemedicine? J. Med. Internet Res. 3(2), e22 (2001), http://dx.doi.org/10.2196/jmir.3.2.e22

[18] Bos, L., et al.: Patient 2.0 Empowerment. In: SWWS 2008, pp. 164–167 (2008), http://science.icmcc.org/2008/07/24/patient-20-empowerment

[19] Becker, E., et al.: A wireless sensor network architecture and its application in an assistive environment. In: PETRA 2008, Athens, Greece, July 15-19, Article No. 25. ACM, New York (2008), http://dx.doi.org/10.1145/1389586.1389616

[20] Loughran, M.: IBM, Public Health Groups Form Global Pandemic Initiative, Goal is to Identify, Map Pandemic Outbreaks and Better Target Vaccines. IBM News (May 2006), http://www-03.ibm.com/press/us/en/pressrelease/19640.wss

[21] Rost-Hein, M., Japs, S.: RFID im Gesundheitswesen, Die Anwendungsbereiche. Informationsforum RFID e.V. (2007), http://www.info-rfid.de/info-rfid/content/e107/e127/e242/rfid_im_gesundheitswesen_ger.pdf

[22] Pohl, K.: Requirements Engineering. Fundamentals, Principles, and Techniques. Springer, London (2010)

[23] Kuehn, A., et al.: Interoperability and Information Brokers in Public Safety: An Approach toward Seamless Emergency Communications. Journal of Theoretical and Applied Electronic Commerce Research 6(1), 43–60 (2011), http://dx.doi.org/10.4067/S0718-18762011000100005

[24] Atzori, L., Iera, A., Morabito, G.: The Internet of Things: A survey. Computer Networks 54(15) 2787-2805 (October 2010), http://dx.doi.org/10.1016/j.comnet.2010.05.010

[25] Wilson, A.D.: Human-Computer Interaction Handbook, Fundamentals, Evolving Technologies and Emerging Applications, pp. 177–199. CRC Press (2007), http://dx.doi.org/10.1201/b11963-10

[26] Van Noordende, G.: Security in the Dutch Electronic Patient Record System. In: SPIMACS 2010, pp. 21–32. ACM, New York (2010), http://dx.doi.org/10.1145/1866914.1866918

[27] Neumann, J., et al.: Integration of audiovisual sensors and technologies in a smart room. Personal and Ubiquitous Computing 13(1), 15–23 (2009), http://dx.doi.org/10.1007/s00779-007-0172-1

[28] Aijaz, F., Chaudhary, A., Walke, B.: Mobile Web Services in Health Care and Sensor Networks. In: ICCSN 2010, pp. 254–259 (2010), http://dx.doi.org/10.1109/ICCSN.2010.42

[29] deBrankart, D.: Swiss ePatient Day – Opening Speech (2011), http://my.brainshark.com/Swiss-ePatient-Day-Opening-Speech-122780377

[30] Wojciechowski, M.: Distributing and Replicating Data in Hospital Information Systems, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.80.4695&rep=rep1&type=pdf

[31] Burkhard, S., Schorre, S.: Ich weiß was du letzten Sommer gelesen hast - E-Reader und die Implikationen für den Datenschutz. In: Schweighofer, E., Kummer, F. (eds.) Europäische Projektkultur als Beitrag zur Rationalisierung des Rechts (Österreichische Computer Gesellschaft), IRIS 2011, pp. 35–44 (2011)

[32] Hevner, et al.: Design Science in Information Systems Research. MIS Quarterly 28(1), 75–105 (2004), http://em.wtu.edu.cn/mis/jxkz/sjkx.pdf

[33] Marion, J.B.: Hospital information systems: perspectives on problems and prospects, 1979 and 2002. International Journal of Medical Informatics 69(2-3) (2002), http://dx.doi.org/10.1016/S1386-5056(02)00098-9

[34] Denzler, L.: Das Arzt-Patienten-Verhältnis im digitalen Zeitalter. Unsere Gesundheitsdaten im Netz, eHealth Publifocus und Elektronisches Patientendossier TA-P10/2008 p. 12 (2008)

[35] Bosshard, G., et al.: Rechtliche Grundlagen im medizinischen Alltag, Ein Leitfaden für die Praxis. SAMW und FMH, Switzerland, pp. 80 (2008)

[36] Breslin, J.G.: Integrating Social Networks and Sensor Networks. In: W3C Workshop on the Future of Social Networking (2009)

[37] Stelle, R.: Social Media, Mobile Devices and Sensors: Categorizing New Techniques for Health Communication. In: ICST Fifth International Conference on Sensing Technology, pp. 187–192 (2011), http://dx.doi.org/10.1109/ICSensT.2011.6136960

[38] Ash, J.S., Berg, M., Coiera, E.: Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors. J. Am. Med. Inform. Assoc., pp. 104–112, http://dx.doi.org/10.1197/jamia.M1471

[39] Bates, D.W.: Using information technology to reduce rates of medication errors in hospitals. BMJ 320(7237), 788–791 (2000)

Older Users' Wish List for Technology Attributes

A Comparison of Household and Medical Technologies

Simon Himmel[1], Martina Ziefle[1], Chantal Lidynia[1], and Andreas Holzinger[2]

[1] Human Computer Interaction Center, RWTH Aachen University, Germany
{himmel,ziefle,lidynia}@comm.rwth-aachen.de
[2] HCI Research Unit, Medical University Graz, Austria
andreas.holzinger@medunigraz.at

Abstract. Facing the increasing user diversity and broad diffusion of technology in work-related and private contexts, the sensible tailoring of technology functionalities, attributes, and interfaces – with reference to the requirements and needs of users – is a key prerequisite of a successful rollout and broad acceptance of technologies. However, user diversity and the specific using contexts of technologies have not been sufficiently researched yet. In this study, we examine the wish list regarding attributes for different technologies in a wide age range. Using qualitative and quantitative methodologies, we explored the different specifications for household and medical devices and assessed which attributes users expect for each of the two different technology types. Exploring user diversity, we analyzed effects of age, gender, and health status on the perception of technology requirements. Results show that not only user diversity but also the specific technology type present as critical factors in the definition of proper attributes of technology. The findings may be useful for human-centered product development.

Keywords: User diversity, technology acceptance, age, gender, purchase and usage criteria, household-, medical technologies, ICT, user expectations.

1 Introduction

During the last decades, human-computer interaction research has made significant gains in understanding technology acceptance of information and communication technology (ICT) and the requirements that need to be considered for a user-centered technology development. Since the pioneering in technology acceptance 25 years ago [1], a huge number of studies corroborated the enormous impact of ease of using a system and its perceived usefulness on the acceptance of a technology (e.g.,[2], [3]). Nevertheless, the knowledge about specific determinants of technology acceptance and the impact of situational aspects of usage contexts is still limited[4]. This is based on the fact that nowadays technology has to cope withmuch more complex using situations compared to the situation of technology usage in former times [5]. A first factor in this context is the user diversity with an ever increasing number of seniors that are confronted with a broad range of technology and urged to understand, learn, and use it

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 16–27, 2013.

[6]-[8]. The second impact is the ongoing diffusion of technical devices. Technology and electronic services are deeply integrated into daily life, thereby raising novel requirements as well as concerns regarding privacy, security, and control [9]. A third factor addresses the different types of technology in different using contexts. One and the same technology, once used in a medical context and once use for daily support, may require completely different using profiles which have not been considered sufficiently within technology acceptance research [10], [11].

For the development of well-accepted future technologies, we need to integrate the users and find out which factors they perceive as relevant for the broad acceptance of different technologies. Technology type, context of use, and age are crucial determinants and highly relevant for the extent of acceptance and the willingness of older people to actually use technology [12], [13].

1.1 Questions Addressed

As users' demands and concerns were neglected within product developmentfor a long time, a considerable knowledge gap exists regarding the needs and requirements of users and their expectations for a useful and functionally sufficient technology. In an exploratory approach, we contrast technologies in different contexts of use, specifically household and medical technologies. Both technology types are very frequently used by older adults, a key targetgroup of modern technology. The questions guiding this research were the following:

(1) *What* are the users' demands for household and medical technologies?
(2) *How* do demands for household and medical technologies differ from each other?
(3) *How* do the demandsvary for different users (gender, age, health status)?

In order to learn about the impact of user diversity, participants of both genders and of a wide age range (19-74 years) were examined.

2 Methodology

The present study was designed to get a deeper understanding of older adults' requirements and expectations of desired functionalities in different technology domains, contrasting household and medical technologies. One of the basic principles of empirically assessing the willingness of technology adoption is the fact that the method used has an effect on the outcome, especially in a participant groupthat is not used to take part in empirical studies (as older adults are) and therefore might be highly receptive to expectations of what seems appropriate in the respective setting.

In order to truly understand older adults' wishes, expectations, and requirements in the different technology domains, we opted for a mix of qualitative and quantitative methodologies. Therefore, we compiled a questionnaire that consisted of both an explorative qualitative part in which participants could freely state their opinions(Part I) and a quantitative part (Part II) in which different aspects had to be evaluated on predefined scales.

2.1 Empirical Research Procedure

In the beginning of the empirical process, demographic data was gathered. Gender, age, and self reported health-status were the independent research variables as they are widely known to be key factors regarding technology adoption and willingness to use technology [3],[4],[5]. Regarding the assessment of their health condition, participants could assign themselves to either a "healthy condition"category or a "not healthy condition" category in case of suffering from chronic disease(s).

Then the qualitative part started (Part I). Participants were given the opportunity to freely state attributes they would require of household and medical technologies respectively. Nine attributes could be given at the most. This qualitative approach was realized in an open format by asking for the users' wish list: Whatattributes should household and medical technologies respectively possess?For both technology types, we operationalized medical and household technologies, giving typical examples from daily experience (e.g. blood pressure meter (medical technology), washing machine (household technology).

Then the quantitative part began (Part II), using the questionnaire, which requested participants to allocate 18 statements either to medical or to household technologies.

For both research parts the dependent variables were the number and nature of attributes ascribed to household vs. medical technologies.

2.2 Participants

In order to study "age" as determining factor for the design of technologies, we examined a wide age range. 36 participants, aged between 19-74 (M=36, SD=18,07), took part in this study. 13 participants were female (36%), 23 male (64%). The sample was split into four age groups (Table 1).

Table 1. Splitting of Participants into Age Groups

Age groups	**N age group**	**% age group**	**% female**	**% male**
Age < 20	6	17	83	17
Age 21-35	17	47	12	88
Age 36-60	8	22	63	27
Age > 60	5	14	20	80

7 participants had at least one chronic disease (19% ill, 81% healthy).Participants were recruited through advertisements in a local newspaper and announcements in public places.Participants were not compensated but volunteered to take part,highly motivated by the fact that they were asked as experts for the design engineering of technology in socially and societally important technology fields.

3 Results

As the present study was mostly exploratory in nature and aimed at uncovering desired functionalities and attributes of technology in different domains, we did not use inference statistical analysis but report the data descriptively (frequency data in %, M=means, SD=standard deviations).

3.1 Part I: Desired Attributes of Household and Medical Technologies

In order to evaluate the qualitative data from the wish list for household and medical technology, the stated attributes were categorized. The categories were based on related research [6],[4] and adjusted in the course of the analysis. The number of attributes mentioned by category can be seen in Table 2.

Table 2. Number and percentage of mentioned attributes:household and medical technologies

	Household Technology		**Medical Technology**	
Categories	$n_{household}$	% household	$n_{med\ tech}$	% med tech
Design	7	4.1	0	0
Other	4	2.3	3	2.3
Price	20	11.6	7	5.5
Documentation	0	0	10	7.8
Facilitation of everyday life	29	16.9	11	8.6
Unobtrusiveness	4	2.3	16	12.5
Functions & usability	58	33.7	21	16.4
Flexibility & mobility	29	16.9	22	17.2
Reliability	21	12.2	38	29.7
Total of attributes	172	100	128	100

As can be seen in Table 2, more answers were given for household technologies, showing that the participants' mental model of attributes ofhousehold technologies is more differentiated than that of medical technologies. Figure 1 depicts the percentage of attributes in the different categories.

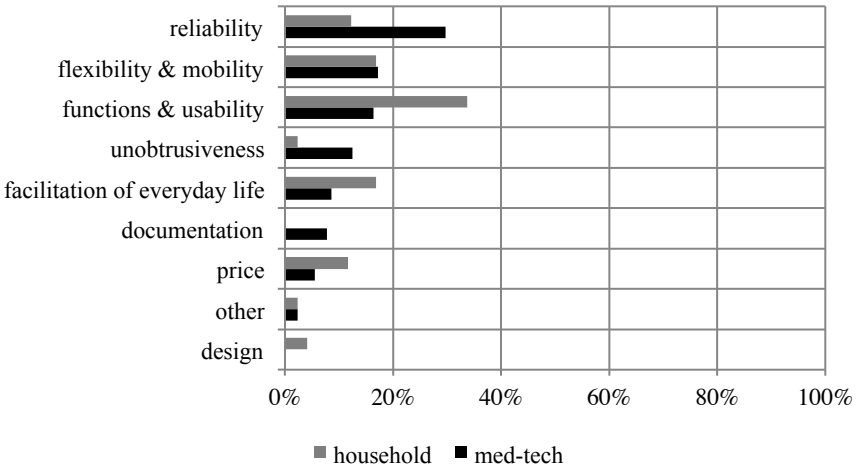


Fig. 1. Mentioned categories for medical and household technologies

As can be seen in Figure 1, desired attributes for the different technology domains differ not only with respect to the mere number of mentioned functionalities.The attribute patterns differed also qualitatively and revealed how very different the wanted attributes are between the two technologies.

"Reliability," for instance, was found to be the key category for medical technologies outdoing all other attributes. Conversely, "functions and usability," obviously less important in case of medical technology, plays the key role in household technologies. Interestingly though, not all attributes depend on the type of technology. "Flexibility and mobility" are similarly wanted in medical and household technologies as the second most important category.

Gender Effect. In this analysis, gender effects are focused on, based onthe question if female and male users require different attributes within the different technology domains. In Figure 2, the requirements for household technologies are depicted.

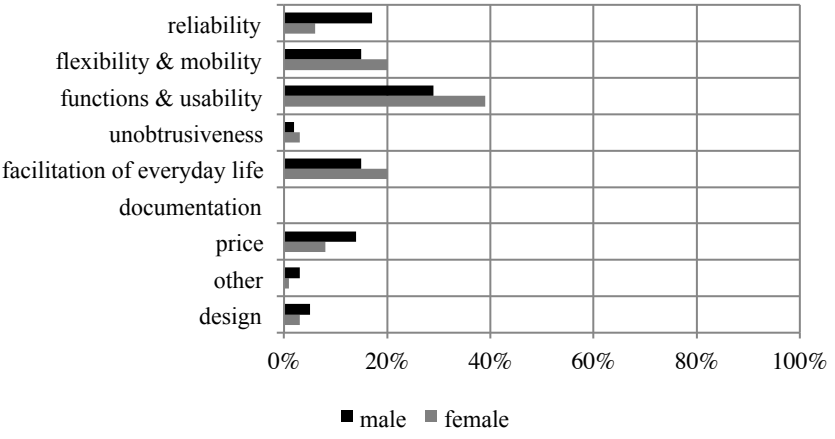


Fig. 2. Mentioned categories for household technologies: gender effects

When looking at household technologies (Figure 2), it is apparent that women attach higher importance to "functions and usability," to "flexibility and mobility" as well as to "facilitation of every day life." Men, in contrast, value "reliability," "price," and "design." Neither gender reported "documentation" requirements, thereby suggesting household technologies areeasy to use and learn, without the use of specific documentation.

In Figure 3, findings of gender-related requirements regarding medical technologies are illustrated. Again, there are considerable gender differences.

The most obvious difference regards the reliability requirement that is more than twice as important for men as it is for women. For women, again "functions and usability," "flexibility and mobility" but also "unobtrusive design" and "facilitation of everyday life" are important key features in the medical technology domain.

Two other findings seem noteworthy in this context. In contrast to household technology, the price is less important in the medical technology domain (more important to men, though) and "documentation" is needed for medical technology devices (more important to women).

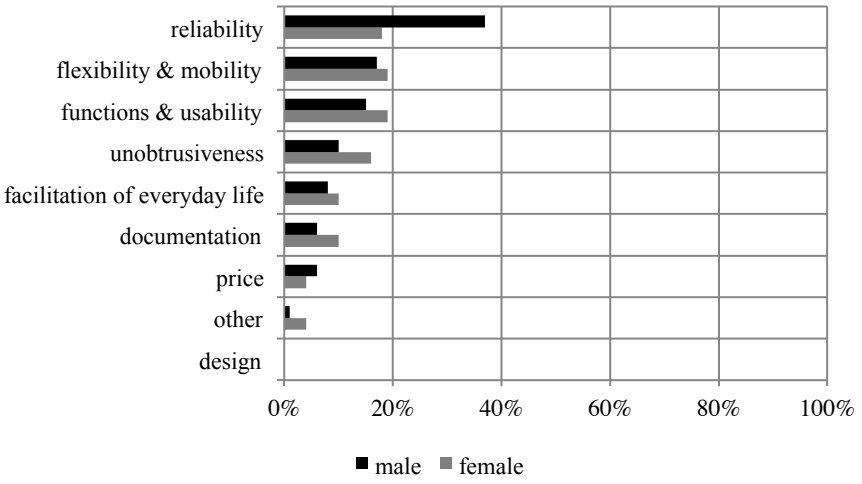


Fig. 3. Mentioned categories for medical technologies: gender effects

Age Effect. When looking for age differences, we see a rather inhomogeneous picture for household technologies (Figure 4).

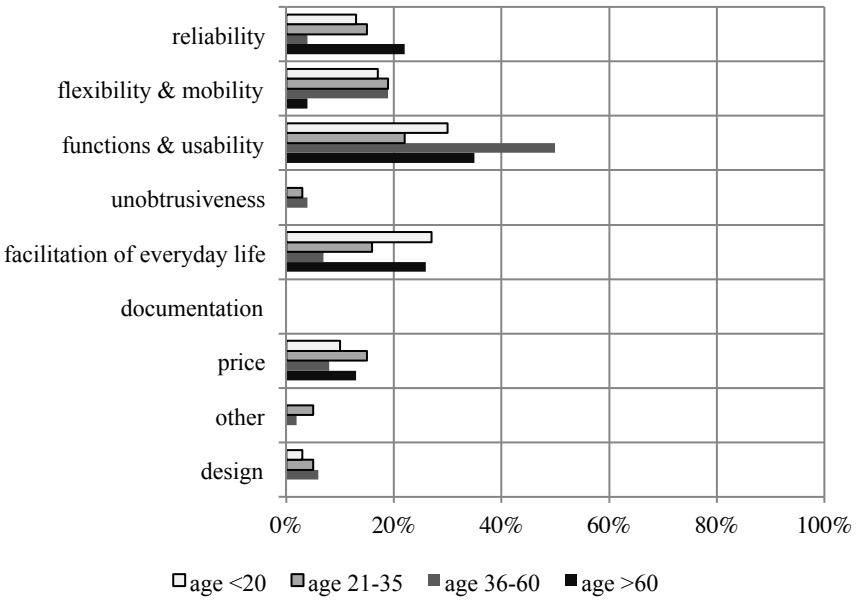


Fig. 4. Mentioned categories for household technologies: age effects

For the oldest group, “functionality and usability,” “facilitation of everyday life,” and “reliability” were the most important features. Across all ages, “functionality and usability,” “facilitation of everyday life,” and “price” seem to be the most important categories for household technologies.

Are age effects also decisive for medical technologies? Figure 5 shows that the requirements are affected to a lesser extent by age. For all age groups, “reliability” of medical technology is the most important feature. “Flexibility” is mentioned as important for all groups except the oldest (over 60 years of age), which did not attach importance to this criterion at all. “Functions and usability” was not as relevant for the oldest group, in contrast to attributes like “unobtrusiveness” and “facilitation of everyday life,” which were found to be crucial characteristics for medical technology in the oldest group.

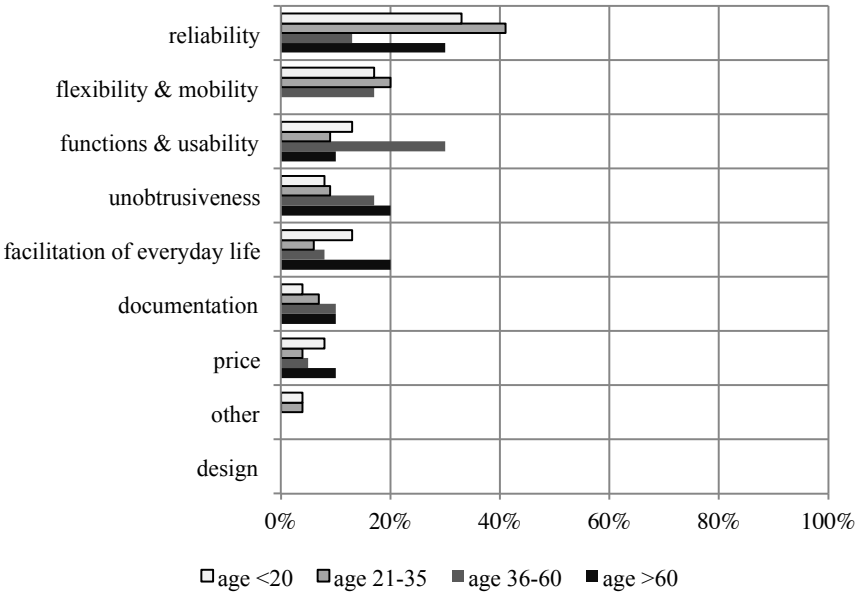


Fig. 5. Mentioned categories for medical technologies: age effects

Effects of Health Status. A final analysis considers the health status of participants. One might expect that the health status should have no effect on the key requirements in household technologies (Figure 6), yet that in medical technologies (Figure 7) healthy persons should have less requirements compared to ill persons,and“reliability,” “flexibility and mobility” but also “usability” should be the key attributes.

Household technologies are considered first (Figure 6). As expected, chronically ill and healthy persons assess the same attributes as (un)important. The most important feature is the usability requirement, independent of the health status. Among the unimportant attributes are “design” and “documentation.”

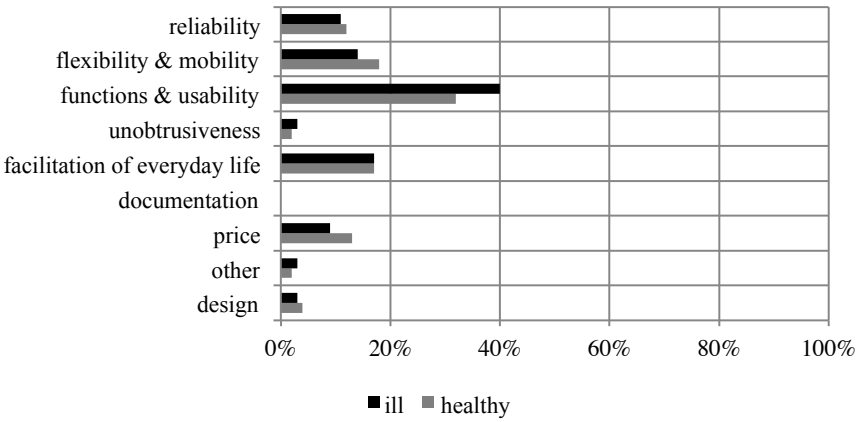


Fig. 6. Mentioned categories for household technologies: effects of health status

When looking at the requirements of medical technology, considerable differences emergedbetween the attributes reported by healthy and ill persons (Figure 7).

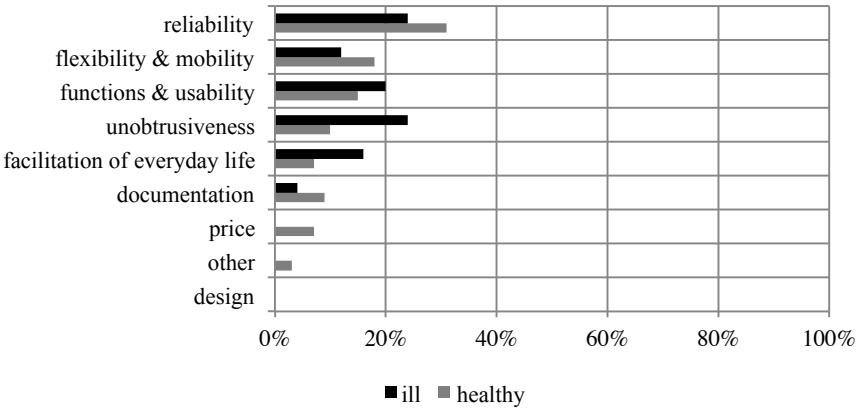


Fig. 7. Mentioned categories for medical technologies: effects of health status

For chronically ill persons, "reliability," "unobtrusiveness," "usability," and "facilitation of everyday life" are most important. While most of the desired functions are easily comprehensible, the requirement for "unobtrusiveness" by ill persons seems to be especially noteworthy. It shows that the stigmatization of illness is a sensitive aspect for peopleand might be compensated by unobtrusive designs. For healthy persons, interesting findings were revealed as well. It is worth mentioning that healthy people attach higher importance to reliability than ill persons do. This finding is contra-intuitive at first sight. Though, possibly, reliability might be less important to ill persons due to their higher domain knowledge of how to handle medical technology. This assumption is corroborated by the higher need for documentation in healthy persons, again referable to their higher insecurity when dealing with disease-related information.

Summarizing the findings so far, we can conclude that users perceive different requirements for technologies in different domains. In addition, the key requirements are also modulated by user diversity. Gender, age, and health status are sensitive factors that need to be considered quite early in the design process of a technology.

3.2 Part 2: Allocated Statements to Household and Medical Technologies

After the qualitative approach, we presented 18 different statements and asked participants to allocate the statements to either medical or household technology.

The selection of the statements was based on interviews carried out prior to this study [7], [8], [9]. The statements represent alternative endings of the sentence "I would use the technology, if," depicting conditional acceptance using motives (Figure 8).

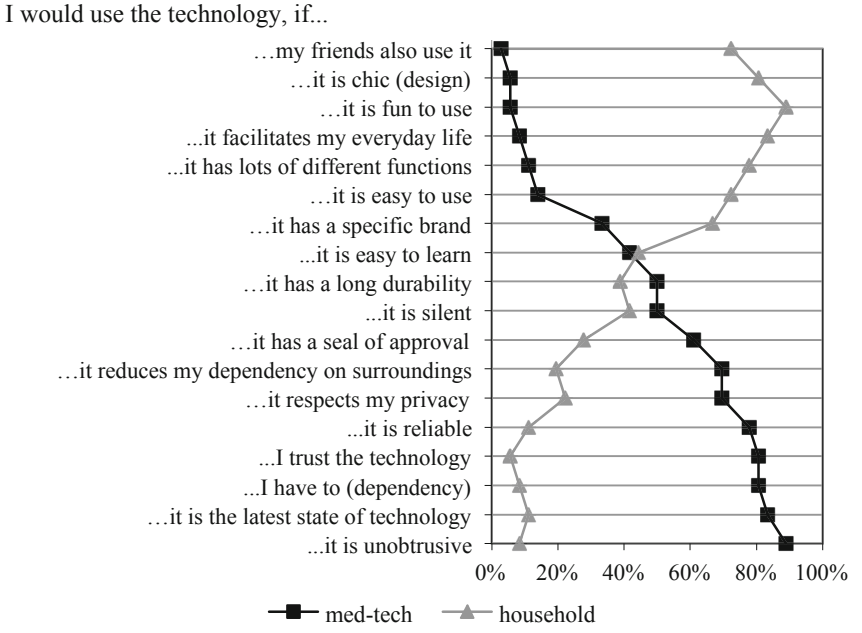


Fig. 8. Conditional acceptance using motives in the different technology fields

It can be seen that there are completely opposing requirement profiles for both technology domains. The curves for medical technology and for household technology cross in one main aspect: the ease of learning the system, followed by a high durability, and a silent mode while using the technology.

When defining the pro-using profile of medical technology, "unobtrusiveness of technical design," "the latest state of technology," "having no alternatives to using it," "trust," and "reliability" are key. The very same arguments are the least important features for household technologies. For the latter, the pro-using profile is characterized by "attractive design," "fun to use," "technology facilitates life," and providing many "different functions." No age, gender,or health effects were revealed, showing that user diversity is indistinctive in the perception of household technologies.

4 Discussion

In this paper we focused on users' wish list regarding attributes of different technology types: household and medical technologies. On the one hand, we identified generic attributes, which should be present in both technologies and independent of user diversity: Here, adequate functionality and high usability are key attributes, but the extent to which technology facilitates everyday life is also important. Furthermore, independent of the participants' age, gender, and health status, there are also some generic attributes that seem to have no importance at all (at least they were not mentioned on the respective wish lists). As such, documentation and instruction come to the fore, which is quite astonishing considering the mostly suboptimal design of many technical devices' manuals. Regarding medical technology, reliability and usability are crucial for all users, which was expected, confirming previous research outcomes in the medical technology context [14], [15]. It is an insightful finding, though, that unobtrusiveness is a critical attribute of medical technology, especially for older and chronically ill persons. One could have expected that unobtrusiveness of technology might be less important for a group that is quite dependent on technology. Apparently, this is not the case and represents the typical designers' error: ignorance towards the customers' experience. As older and ill participants noted, the stigmatization of being old and ill is very negatively biased in our society and therefore unobtrusiveness is indispensable for people using the technology.

Regarding the methodology, the mix of qualitative and quantitative procedures revealed a rich field of insights and allowed us to screen user diversity and technology contexts in a very sensible way. The task of providing a "wish list" for different technology types that are very familiar to the participants pleased them. In addition, participants were highly motivated to work on technology design principles and were very willing to contribute their knowledge. They appreciated being integrated into technology development as key users, especially as this opposes the traditional designers' approach of first producing a technology and later realizing said technology designs might have severe usability barriers or are outright rejected by the (intended) target group.

A final remark is directed to the older persons as an important future user group. In contrast to the usual procedure in which younger technical designers develop technology for older users by just imagining what could be useful or necessary for senior persons. We should be aware that this procedure is naïve if not ignorant. Aging is complex and quite differential [16], [17]. Not all users age in the same way and the requirements, the needs and whishes towards a well-accepted technology might be individual. In addition, designers should be aware of the fact that aging and technology generation should be distinguished. Even though the age-related decrease of sensory and motor abilities might be comparable over times, the acceptance for and the requirement towards humane technology might be different in the different generations, reflecting upbringing aspects, mental model of how technology works and different societal needs [17]. Thus, including users into early stages of technology developments is indispensable in order to reach broad acceptance of technology [18], [19].

5 Limitations and Future Research Duties

Finally, there are some limitations that will have to be addressed in future work. One is the comparatively small sample size. Even though qualitative research approaches often have small sample sizes, we cannot expect a broad generalization of the findings unless we validated the outcomes with a larger sample size. This will be accomplished in future work. Also we are aware that household and medical technology are only exemplary technology fields, which might be relevant against the background of the demographic change. As a matter of fact there are other technology developments, smart clothing or robot development, which should be also examined in this context. The other is the focus on a decidedly European perspective. In future studies, we will have to work out to what extent insights won here also apply to other cultural backgrounds and societies with different value systems, different education levels, and different economic structures[20], [21].

Acknowledgements. Authors thank participants, especially the older ones for their time and patience to volunteer in this study and to allow insights into a sensitive topic. Thanks also to Vanessa Schwittay for research support.

References

[1] Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 319–340 (1989)
[2] Arning, K., Ziefle, M.: Different perspectives on technology acceptance: The role of technology type and age. HCI and Usability for e-Inclusion, 20–41 (2009)
[3] Wilkowska, W., Ziefle, M.: Which Factors Form Older Adults' Acceptance of Mobile Information and Communication Technologies? In: Holzinger, A., Miesenberger, K. (eds.) USAB 2009. LNCS, vol. 5889, pp. 81–101. Springer, Heidelberg (2009)
[4] Wilkowska, W., Ziefle, M.: User diversity as a challenge for the integration of medical technology into future home environments. In: Human-Centred Design of eHealth Technologies. Concepts, Methods and Applications, pp. 95–126 (2011)
[5] Rogers, Y.: The Changing Face of Human-Computer Interaction in the Age of Ubiquitous Computing. In: Holzinger, A., Miesenberger, K. (eds.) USAB 2009. LNCS, vol. 5889, pp. 1–19. Springer, Heidelberg (2009)
[6] Mitzner, T.L., Boron, J.B., Fausset, J.B., Adams, A.E., Charness, N., Czaja, S.J., Dijkstra, K., Fisk, A.D., Rogers, W.A., Sharit, J.: Older adults talk technology: Technology usage and attitudes. Computers in Human Behavior 26(6), 1710–1721 (2010)
[7] Tsipi Heart, E.K.: Older adults: Are they ready to adopt health-related ICT? International Journal of Medical Informatics (2011)
[8] Kleinberger, T., Becker, M., Ras, E., Holzinger, A., Müller, P.: Ambient intelligence in assisted living: enable elderly people to handle future interfaces. Universal access in human-computer interaction. Ambient Interaction, 103–112 (2007)
[9] Wilkowska, W., Ziefle, M.: Privacy and data security in E-health: Requirements from the user's perspective. Health Informatics Journal 18(3), 191–201 (2012)
[10] Ziefle, M., Himmel, S., Holzinger, A.: How Usage Context Shapes Evaluation and Adoption in Different Technologies. Advances in Usability Evaluation 21, 211 (2012)

[11] Falaleeva, N.G., Johnson, R.D.: Influence of individual psychological traits on attribution toward computing technology. In: Eighth Americas Conference on Information Systems vol.1028-1033 (2002)
[12] Brown, S.L., Eisenhardt, K.M.: Product development: Past research, present findings, and future directions. Academy of Management Review, 343–378 (1995)
[13] Nagamachi, M.: Kansei Engineering: A new ergonomic consumer-oriented technology for product development. International Journal of Industrial Ergonomics 15(1), 3–11 (1995)
[14] Ziefle, M., Himmel, S., Wilkowska, W.: When your living space knows what you do: Acceptance of medical home monitoring by different technologies. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 607–624. Springer, Heidelberg (2011)
[15] Ziefle, M., Wilkowska, W.: Technology acceptability for medical assistance. In: 4th ICST Conference on Pervasive Computing Technologies for Healthcare, pp. 1–9 (2010)
[16] Ziefle, M., Schaar, A.K.: Technical Expertise and its Influence on the Acceptance of Future Medical Technologies. What is influencing what to which extent? In: Leitner, G., Hitz, M., Holzinger, A. (eds.) USAB 2010. LNCS, vol. 6389, pp. 513–529. Springer, Heidelberg (2010)
[17] Ziefle, M., Jakobs, E.-M.: New challenges in Human Computer Interaction: Strategic Directions and Interdisciplinary Trends. In: 4th International Conference on Competitive Manufacturing Technologies, pp. 389–398. University of Stellenbosch, South Africa (2010)
[18] Ziefle, M.: Modelling mobile devices for the elderly. In: Khalid, H., Hedge, A., Ahram, Ż. (eds.) Advances in Ergonomics Modeling and Usability Evaluation, pp. 280–290. CRC Press, Boca Raton (2010)
[19] Arning, K., Ziefle, M.: Ask and you will receive: Training older adults to use a PDA in an active learning environment. International Journal of Mobile Human-Computer Interaction 2(1), 21–47 (2010)
[20] Sensales, G., Greenfield, P.M.: Attitudes toward Computers, Science, and Technology A Cross-Cultural Comparison between Students in Rome and Los Angeles. Journal of Cross-Cultural Psychology 26(3), 229–242 (1995)
[21] Alagöz, F., Ziefle, M., Wilkowska, W., Valdez, A.C.: Openness to Accept Medical Technology - A Cultural View. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 151–170. Springer, Heidelberg (2011)

Evaluating the Energy Efficiency of OLTP Operations

A Case Study on PostgreSQL

Raik Niemann[1,2], Nikolaos Korfiatis[2], Roberto Zicari[2], and Richard Göbel[1]

[1] Institute of Information Systems, University of Applied Science Hof, Hof, Germany
[2] Chair for Database and Information Systems, Institute for Informatics and Mathematics, Goethe University Frankfurt, Frankfurt am Main, Germany

Abstract. With the continuous increase of online services as well as energy costs, energy consumption becomes a significant cost factor for the evaluation of data center operations. A significant contributor to that is the performance of database servers which are found to constitute the backbone of online services. From a software approach, while a set of novel data management technologies appear in the market e.g. key-value based or in-memory databases, classic relational database management systems (RDBMS) are still widely used. In addition from a hardware perspective, the majority of database servers is still using standard magnetic hard drives (HDDs) instead of solid state drives (SSDs) due to lower cost of storage per gigabyte, disregarding the performance boost that might be given due to high cost.

In this study we focus on a software based assessment of the energy consumption of a database server by running three different and complete database workloads namely TCP-H, Star Schema Benchmark -SSB as well a modified benchmark we have derived for this study called W22. We profile the energy distribution among the most important server components and by using different resource allocation we assess the energy consumption of a typical open source RDBMS (*PostgreSQL*) on a standard server in relation with its performance (measured by query time).

Results confirm the well-known fact that even for complete workloads, optimization of the RDBMS results to lower energy consumption.

1 Introduction

A general assumption is that efficient programs are also energy efficient [8]. This seems to be reasonable since a program requiring less computation time will probably also require less energy for its execution. This can be true if we assume a constant energy use of a computer, and the computer performs more tasks in the same amount of time because of time efficient programs.

However, a more detailed analysis may show a different picture. For example it is a well-known fact that one complexity measure like time can be optimized at the costs of another complexity measure of space. An optimization following this path may require that the increased amount of data for a program may be

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 28–43, 2013.

transferred from one type of memory to a different type of memory. In a modern computer this might result that the data may be moved from the CPU cache to the main memory or from main memory to a drive. An obvious consequence of this transfer are longer access times which may already reduce the benefit of these optimization strategies. In addition the usage of a different memory type may result in higher energy costs [14]. For example, this can be the case if a large data structure needs to be generated on an external drive and the access to this drive would need to be frequently used in comparison to a solution where the drive may be even stopped to save energy.

Since the amount of information to be processed has drastically increased over the last couple of years, large and very large data centers which are expected to support internet operation have been built by major players (e.g. *Google* and *Amazon*). However, increased costs of operation (electricity, cooling costs) have made these vendors to offer the rental of unneeded data center capacity (disk storage space, computation time) to nearly everyone who is willing to pay. This gives rise to the concept of "*cloud computing*" where, for example, server capacities are only needed for a limited time or where the acquisition of an own data center is too expensive [2].

Furthermore, in the context of database applications another issue might arise: the distribution of data across a database cluster as a basis for distributed computing, as it happens in the case of Big Data/*Hadoop* clusters [1]. Industry trends[1] advocate that increased demand for server performance will be fulfilled by deploying more (database) servers and data centers. But with increased energy costs, data centers operators are looking for mechanisms to avoid or reduce those costs (scale-up strategies). This way the energy efficiency of a single database server becomes important as it can be a crucial component for the overall cost assessment of a data center operation where energy costs are the most significant factor on their operation and scalability.

1.1 Background and Motivation

Several approaches which can be found in the literature have addressed the issue of energy efficiency of database servers from different aspects. *Harizopoulos et. al* [13] as well as *Graefe* [5] divided energy efficiency improvements into a hardware and a software part. Recent work such as the one by *Baroso* and *Hölzle* [3] has revealed the fact that the impact of hardware improvements on the energy efficiency is quite low. A main approach in that direction is the use of Solid State Drives (SSD's) which have been shown to improve performance on typical database operations such as sorting and therefore energy efficiency [4]. Obviously SSDs are faster in such kind of scenarios [10], but there are other tradeoffs as a replacement for HDDs with most obvious the cost per gigabyte (*Schall et. all* [11], *Schröder-Preikschat et. all* [12], *Härder et. all* [6]).

On the other side, software improvements seem to be more attractive when it comes to increasing the energy efficiency of a database server, especially for

[1] *State of the data center 2011* by *Emerson*.

relational databases. *Lang* and *Patel* [16], for example, proposed a database query reordering technique to influence the energy consumption of the database server. *Xu et. all* [20] proposed a modification of the query planner in order to take the estimated energy consumption into consideration.

Tsirogiannis et. all [14] had a very detailed investigation regarding the relation between performance and energy consumption in a RDBMS. For that they did extensive measurements with various hardware configurations that are typically found in a database server for a scale-out scenario. In addition to this, they identified performance tradeoffs for different SQL query operators.

1.2 Objective of This Study

This study provides the basis for a more detailed analysis of the energy consumption in the context of database applications that are common in daily workloads of enterprise users. For this purpose the paper does not only consider the total power consumption of the full system but also the usage of energy by different components. In addition the paper analyses standard energy saving strategies coming with a modern computer system and their impact on the power consumption.

In the context of database applications, it is important to know which database operation or which combination of SQL operators affects which of the measured components. This study analyzes these combinations and their fraction of the overall power consumption.

All the mentioned aspects lead to the final view on the energy efficiency of a single database server. This is important when it comes to a scale-out scenario that is typically found in a database cluster. There are two choices: either one decreases the energy consumption of the used hardware equipment with a small decrease of the database response time, or increases the performance while accepting a slight increase of the energy consumption. Both ways improve the energy efficiency.

As suggested by *Xu* [18], more experimental research has to be done concerning the energy consumption of database servers. Taking this into account, this study tries to have a closer look on both choices mentioned above and to give recommendations how to improve the energy efficiency of a single database server in a general way (for example the usage of buffers and indices combined with traditional HDDs as the primary storage for the database files).

2 Measurement Methodology

2.1 Test Server Preparation

For the purpose of this study, we constructed a database server making use of recent technical core components. The operating system selected for testing was a typical *Linux* distribution (*Ubuntu Server* version 11.10) using a stable kernel (kernel version: 3.0.0.17). In order to eliminate biases from the operating system in our

Table 1. Hardware characteristics of the test server used in our study

CPU	*Intel Core i7-860 @2.8 GHz*
Main memory	3x *Samsung* DDR-2 2 GByte 800 MHz (m378b5673fh0-ch9)
Hard drives	3x *Hitachi* 1 TByte, 16 MByte cache (HDT721010SLA360)

measurements, all unnecessary operating system services were turned off. The database management system (DBMS) that was used was *PostgreSQL* version 9.1. The hardware characteristics of the test server are provided in Table 1.

Two of the three hard drives were combined as a striping RAID array in order to boost performance and to separate the filesystem calls of the DBMS from the operating system. The operating system itself was installed on the remaining third hard drive.

We identified the core components we were interested in their power consumption throughout the various tests as follows: (a) CPU, (b) main memory, (c) motherboard as a whole and (d) the hard drives of the RAID array. These core components were used as a unity of analysis in order to assess the optimization options. Additionally we were interested in the impact of the operating system settings as well as the test server capabilities on the energy consumption of the core components. Modern server configurations don't allow an external measurement of the energy consumption of internal server components so we decided to use a test apparatus based on moderate hardware. Our test arrangement makes use of a modified ATX power for measurements but server manufactors mostly use proprietary power cords. The complete test arrangement can be viewed online[2].

Besides, the test server's motherboard used *Intel*'s EIST[3]. In general *EIST* enables and disables CPU cores or reduces and raises the overall CPU clock frequency as a function of the CPU usage. This also affects other technical aspects, e.g. the heat dissipation from the CPU.

Recent Linux kernels offer several modules providing more data to the frequency scaling heuristics. The observed modules for our test server configuration are the *on-demand* and the *power-saving* modules. Taking this into account, the command line tool *powertop*[4] was used to suggest configurations on the disablement of operating system services that might influence energy consumption.

2.2 Stress Tests on Energy Consumption

Before assessing the energy efficiency as a whole as well as the components we were interested in profiling their energy consumption, we performed stress tests to get a detailed overview of the energy distribution among the single

[2] Refer to `http://team.iisys.de/test-apparatus/`

[3] Acrynym for *Intel Enhanced Speedstep*. It allows the system to dynamically adjust the CPU frequency and voltage to decrease energy consumption and heat production.

[4] Refer to `http://www.lesswatts.org/projects/powertop/`

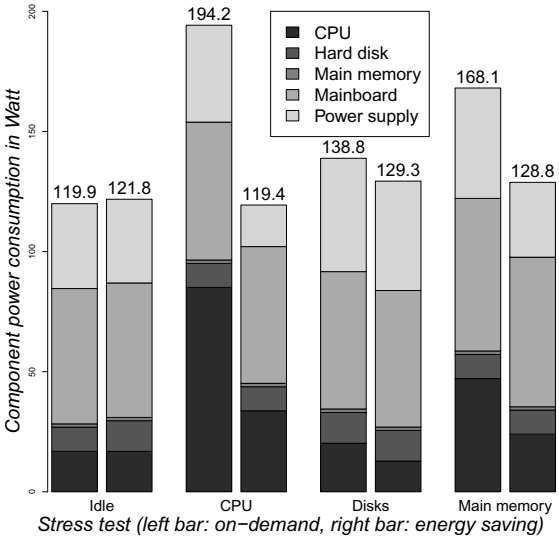


Fig. 1. Energy consumption per component during our calibration stress test

components. Figure 1 summarizes the stress tests. Each tuple of the shown bars represents a stress test for a core component. The left bar of a specific tuple illustrates the energy consumption of each component in conjunction with the *on-demand* frequency scaling module. The right bar of the tuple displays the energy consumption of all energy saving settings enabled, respectively. Figure 1 also shows the energy consumption of the power supply unit (PSU) which is the difference between the overall power consumption of the test server and the one of the measured components.

Our first action was to calibrate the test procedure by analyzing the energy consumption of the core hardware components as a basis for the other tests. All services and applications not required for the operating system as well as the DBMS service were turned off. The operating system uses default settings, e.g. the usage of the *on-demand* CPU frequency scaling module. The average consumption per core component is shown in Fig. 1 in the left bar of the tuple named *Idle*. The right bar of this tuple shows the energy consumption with all available energy saving settings turned on. There is a slight increase of the power consumption: although the CPU is forced to reach the energy saving idle states, it is constantly interrupted by doing so, for example by administrative background processes of the operating system. Changing the CPU state costs some effort but results in higher energy consumption.

To measure the power consumption stressing the CPU of our test server, we decided to use the *Linux* command line tool *burnMMX* because it utilizes all parts of the CPU including extension like *MMX* or *AVX*. Based on the eight CPU cores reported to the operating system, we ran our CPU stress test eight times and increased the number of running *burnMMX* instances accordingly. As assumed the power consumption increases up to the four real cores and

remains relatively constant if more cores were used due to HT[5]. This behavior corresponds to the CPU tests reported in [14]. The CPU stress test with the highest energy consumption for both scenarios (energy saving settings turned off and on) is depicted in Fig. 1 as the bar tuple labeled *CPU* for comparison with the other stress tests.

To stress test the hard drives in the RAID array we executed the I/O benchmark suite *iozone* while observing the power consumption throughout the different benchmark tests (different access strategies, buffered and un-buffered data access and so on). Our test results are displayed in Fig. 1 in the second bar tupel labeled *Disks*. Our test shows that the impact on the energy consumption is low when all energy savings settings are turned on. However, in contrast to the power consumption of current SSDs [12] the power consumption of the hard drives is two to three times higher.

Finally we stressed the main memory with the command line tool *memtester* that uses different patterns to access the main memory. It also tests for the correctness of the contents by writing, reading and comparing the main memory areas. The effect on the power lane supplying the main memory was not measurable. In contrast to this, the left bar of the tuple named *Main memory* in Fig. 1 indicates an increase of the overall power consumption of 30 Watts in which the CPU is responsible for. Even with all energy saving settings turned on [19], the overall energy consumption was higher than in idle state.

2.3 Measuring Energy Efficiency

We define the performance ratio (P) of a single database query as the unit of time to execute the query in time (t) to obtain the results [15]:

$$P = \frac{1}{t} \tag{1}$$

We then normalize a set of performance values to values between zero and one as follows:

$$P_{normalized,i} = P_i \cdot \frac{1}{max(P)}$$

We then define the *energy efficiency* EE as the ratio between the performance P of the database query and the electrical work W executing the query:

$$EE = \frac{P}{W} \tag{2}$$

We then normalize the set of efficiency values to values between zero and one as follows:

$$EE_{normalized,i} = EE_i \cdot \frac{1}{max(EE)}$$

[5] *HT* is an acronym for *Hyperthreading* by *Intel*. It is used to improve parallelization of computation.

2.4 Selection of Workloads and DBMS Parameters

As aforementioned our intuition here is to examine the effects of executing a database query in relation to the overall power consumption. For example a combination of a not well formed SQL query and an optimized query planner can cause a cascade of operations that consumes a lot of unnecessary power, e.g. sequential scans cause unnecessary hard drive accesses with all its overhead in the operating system (access control, swapping and so on).

According to related work, for example [14] or [3], we identified several settings for our *PostgreSQL* database server we hypothesized they had an important impact on the power consumption. Those settings can be divided into two groups: the first one are the settings for the underlying operating system and for *PostgreSQL* and the second one are settings for the database itself. In detail those settings are:

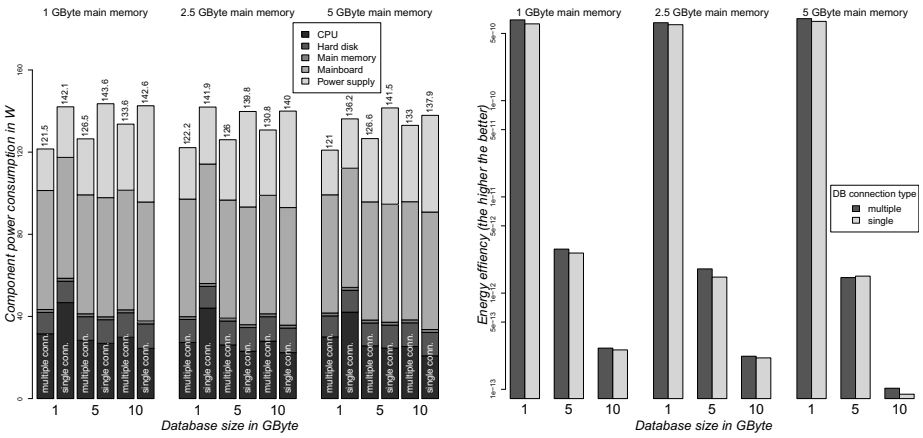
- The size of the main memory assigned to *PostgreSQL* to operate
- The size of the miscellaneous buffers, e.g. for sorting resulting rows or caching
- The settings for the query planner
- The size of the data in the database
- The session type of executing subsequent queries (single session vs. multi session)
- Combinations of the different SQL operators and functions, for example increasing number of joined tables or the number of result set dimensions to be restricted

To get comparable results we decided to run three complete database benchmark workloads: *TPC-H* [9], the *Star schema benchmark* (SSB [7]) and a third benchmark constructed specifically for this study, we call it the *W22* benchmark. The first two offer a standardized way for comparison whereas the last one is a workload we composed to analyze the behavior of *PostgreSQL* not covered by the previously mentioned workloads.

3 Workload Results

3.1 TPC-H Workload

For the TPC-H workload we used their data generator to generate three databases with a size of 1, 5 and 10 GByte of data as well as all suggested indices. We chose these database sizes because we wanted the databases to fit completely, nearly and under no circumstances in the working main memory for *PostgreSQL* by assigning 1, 2.5 and 5 GByte. This selection was made due to the limitation of 6 GByte overall main memory present in the database server. In a final setup step we optimized the internal data structures and the query planner statistics of *PostgreSQL* for the created databases.



(a) Energy consumption per component (b) Energy efficiency

Fig. 2. Energy consumption per component and energy efficiency for the TPC-H workload

At first we ran the benchmark queries subsequently using a new database connection to avoid *PostgreSQL*'s internal cache[6]. After this we retried the TPC-H benchmark using a single database connection for the SQL queries. The average power consumption per component for both test series is shown in Fig. 2(a).

This figure clearly indicates the higher energy consumption when a single session is used. We recognized that the average overall power consumption for all of our TPC-H tests does not vary a lot. Please compare only the bars which are identically labeled with each other. We assumed that the CPU and hard drives are the main energy consumer but in fact it turned out that the mainboard is the biggest one.

We also assumed that the usage of single database connection would improve the performance because of the better usage of *PostgreSQL*'s internal cache but the opposite occured: the performance was nearly the same with an increased power consumption of 7 percent on average. Using the equations outlined in Sec. 2.3, this leads to a lower energy efficiency of nearly 8 percent on average as depicted in Fig. 2(b).

A detailed study of the query plans reveals a broad usage of sequential scans on the TPC-H database tables as well as a low usage of the internal caches. Therefore we modified *PostgreSQL*'s settings regarding the query planner and caches to favor the provided indices and repeated the tests. This has an inverted effect: the performance decreases by about 4 percent on average. The reason is the overhead to process the indices which are also disk bounded.

[6] Notice that *PostgreSQL*'s implementation isolates client connections completely by running the client sessions in shared-nothing processes. The client processes only share some IPC memory for synchronization purposes and the ACID functionality. This means that every client has its own cache.

In general we observed a big effect on the energy efficiency by assigning a higher portion of the main memory to *PostgreSQL*. This results in massive swapping for some TPC-H queries, e.g. 1, 9 and 21, and the suspension of the *PostgreSQL* process. In fact, *PostgreSQL* is very disk bound and by accessing the database files the operating system swaps heavily because of the reduced main memory portion. Besides, swapping does effect only the hard drive and not the CPU. This affects the overall power consumption for our TPC-H tests and explains the small variation of the values.

Finally we were interested in performance versus energy efficiency ratio as stated in the conclusion of [14]. This ratio for our entire TPC-H tests with all its different configurations is shown in Fig. 3 and confirms the strong relation between performance and energy efficiency (*the most energy-effiencent configuration is typically the highest performing one*).

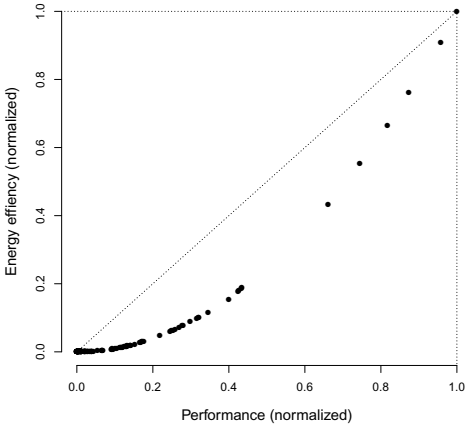


Fig. 3. Performance vs. energy effiency for TPC-H workload

This figure also shows that the majority of the configurations are clustered in the lower left area. This means that most of the queries of the TPC-H benchmark show a poor energy efficiency.

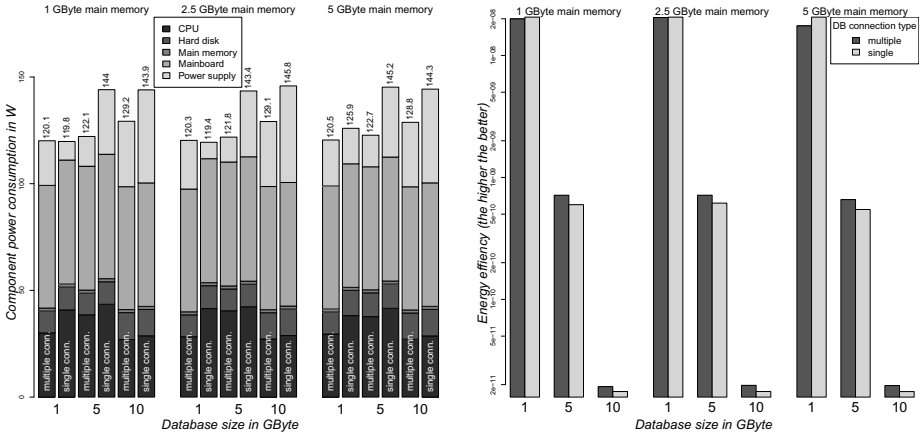
3.2 Star Schema Workload

The *Star schema benchmark* (SSB) was initially composed to get a database layout closer to reality compared to TPC-H which it is based on. According to [7] several original tables were decoupled to make many join operations unnecessary and the set of SQL queries of SSB were created to be more realistic and to test the database capabilities regarding range coverage and indice usage.

For executing the SSB SQL queries we used the same parameters as described in Sec. 3.1: we generated three different databases with 1, 5 and 10 GByte of data and created test configurations for *PostgreSQL* with 1, 2.5 and 5 GByte main memory to work on.

Just as for the TPC-H workload, we ran the SSB database queries subsequently by using a single and multiple database connections. The average energy consumption per component is depicted in Fig. 4(a). The type of accessing the database matters: although the average performance is nearly the same, the average power consumption for multiple connections is lower than the one for using a single connection. The latter consumed nearly 10 percent more energy on average. Except for the database size of 1 GByte, this results in a lower energy efficiency.

In general and in terms of energy efficiency, the SSB benchmark performs better than TPC-H. In addition to this, SSB organizes its SQL queries into groups in which the requested rows do not overlap and the group number indicates on how many dimensions the result set has to be restricted. This allows a better analysis of the results and avoids side effects interfering the results. Figure 4(b) shows the energy efficiency of our SSB tests.



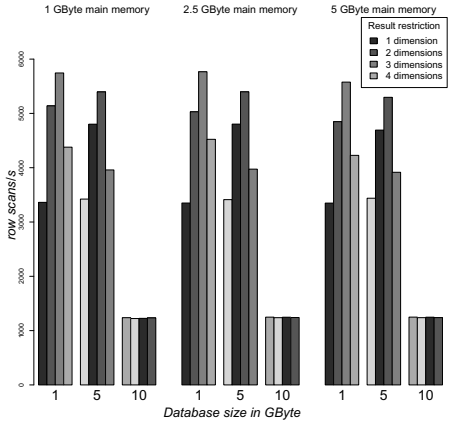
(a) Energy consumption per component

(b) Energy efficiency

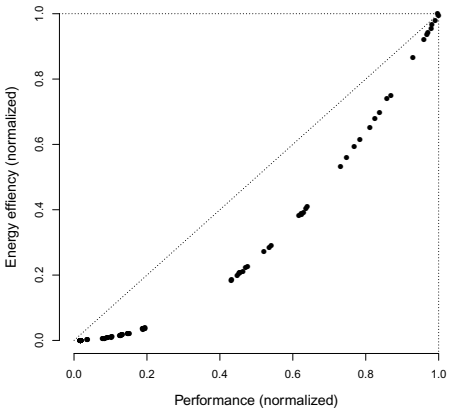
Fig. 4. Energy consumption per component and energy efficiency for SSB workload

A deeper investigation of the query plans showed us that the reduced set of tables and the modified queries lead to a more stable and predictable behaviour concerning the power consumption and energy efficiency. In particular this can be seen in Fig. 4(b): the energy efficiency for a given constant SSB database size does not vary a lot and is relatively independent from the amount of main memory spent for *PostgreSQL*.

Besides, the energy efficiency of the SSB database with 10 GByte is remarkably lower than the ones with 1 and 5 GByte of data. An investigation of the logged system activities during the tests revealed heavy swapping actions which caused the suspension of the *PostgreSQL* process handling the queries. As illustrated in Fig. 4(a), the low CPU energy consumption indicates the heavy swapping activity.



(a) SSB row scans per second



(b) Performance vs. energy efficiency for SSB

Fig. 5. Row scan count and performance vs. energy efficiency for SSB

The query plans also show the general use of sequential scans on the SSB database tables. In fact, no indices were involved to execute the queries in all configurations. The execution of the queries purely relies on the read performance of the database files. As Fig. 5(a) indicates, it does not matter if the result set of the joint tables are further restricted. In addition to this, this figure shows that main memory spent for buffers is relatively unimportant: if one consider a specific database size, the row scan per second rate does not vary a lot. This means that the mentioned rate is independent of the main memory fraction assigned to *PostgreSQL* and also independent of the number of dimensions the result set is restricted.

The performance vs. energy efficiency ratio for all SSB test configurations is illustrated in Fig. 5(b). The same strong relationship can be seen as in Fig. 3 for the TPC-H benchmark tests. In contrast to TPC-H, the vast majority of the SSB test configurations is clustered in the upper right corner of Fig. 5(b). This means that the SSB configurations perform much better. This leads to a better energy efficiency.

3.3 W22 Workload

To get a more in-detailed look at the impact of the different SQL operations and their combinations we composed a database workload called W22. The workload consists of six groups of database queries:

1. aggregate functions (`count`, `avg` and `sum`)
2. grouping (`GROUP BY`)
3. sorting (`ORDERED BY`)
4. selecting different data types, e.g. `int`, `varchar`, `text` and `date`
5. removing duplicates (`DISTINCT`)
6. joins (cross joins, conditional joins)

In contrast to this, the TPC-H and the SSB workload are designed for benchmark an OLAP scenario. Their database queries use a mix of different SQL operations coming from the groups above. Our W22 workload instead tries to analyze the impact of SQL queries in each group. Besides, this workload allows to analyze the behaviour of *PostgreSQL*'s internal query optimizer and query planner as well as the interaction with the underlaying operating system. The W22 queries operate on the TPC-H databases described in Sec. 3.1.

Aggregate Functions. Our first tests with aggregate functions, e.g. `avg()`, `sum()` and `count()`, show in particular the impact of the filesystem cache of the underlaying operating system.

All of the mentioned aggregate functions cause *PostgreSQL* to perform a sequential scan. The first query that was executed (`count(*)`) had a notably longer execution time compared to the next ones (`avg()` and `sum()`). The query plans for all three queries are the same. After a closer investigation of the logged activities of the operating system, we identified the filesystem cache as the performance booster by caching the database files that *PostgreSQL* uses. In this case the kind of executing the queries (single session vs. multi session) does not matter. The most important setting is the fraction of main memory assigned to *PostgreSQL*: the lower the fraction the more main memory is available for system caches.

The suggested optimizations of the queries, e.g. the use of an indexed column for `count()` (`count(column)` instead of `count(*)`), did not change the execution times.

Grouping and Sorting. The queries of group 2 (grouping) and group 3 (sorting) expands the `count()` query from group 1 with SQL operators like `ORDER BY` and `GROUP BY` to get comparable results. Those queries were composed to study the impact of the mentioned SQL operators in reaction of the previous tested benchmarks TCP-H and SSB.

The tests result indicates only a slight difference in terms of execution times and energy consumption compared to the test results of group 1. The impact on the energy consumption of the CPU and the main memory was not measurable. As a result the calculated energy efficiency remains the same.

Selection of Different Data Types. According to our tests, the selection of a specific data type, e.g. `VARCHAR`, `DATE` and `INTEGER`, has no impact on the query execution time.

The other purpose of the queries of group 4 was to test the operators of the different data types with and without having an index on the particular table column. We observed a big impact on the execution time when an index was used whereas the average energy consumption was slightly higher. The presence of an index in combination of an operator[7] supporting this index lead to an immense

[7] For example, the operators greater than, less than and equals (`<`, `>` and `=`) are valid for B-tree indices on numeric table columns in *PostgreSQL*. There are other index types, e.g. inverted index, for other column data types as well as their specialized operators.

performance gain. Therefore the energy efficiency is quite high compared to a sequential scan. In contrast to the queries with an absent index, the queries with an involved index showed an almost linear performance. This is crucial since we used the largest table of TPC-H, `lineitem`, which also has the greatest amount of rows throughout all queries of this group.

PostgreSQL uses an index for a query when a) the index supports the operator of the query and b) the costs for processing the index are lower than sequentially scanning the table. Those costs are composed of customizable base costs and dynamic cost estimations that *PostgreSQL* gathers periodically and statistically from all tables of a database.

We modified the settings for the base costs of loading and processing a database and an index page (usually 8 KBytes of data) to favor the usage of indices, e.g. if the database and indice files are stored on different storage devices with different access speeds (the indice files are usually stored on the faster one). Our test results remain the same because the data and index files are stored on the same hard drive device.

Joining Tables and Eleminating Duplicates. Based on our TPC-H and SSB benchmark results, we were interested in the behaviour of *PostgreSQL* dealing with table joins. There are two kinds of joins supported in *PostgreSQL*: unconditional and conditional joins.

Our first query of this W22 group deals with an unconditional join of two tables where the cross product is further restricted by the conditions given after the `WHERE` clause (one restriction per involved table). We expected this query to be unperformant due to the cross product and the successive restriction of the result set, but this was not the case. The query plan reveals the (unintentional) use of an index for one restriction and a sequential scan for the other one. So the results (performance, energy consumption and the energy efficiency) are the same as mentioned in the last section although a sequence scan is part of the query plan.

The other queries of this W22 group were composed to join two TPC-H tables using inner[8] and equi-joins[9] to examine differences in the query plans. Interestingly those queries indicate the same characteristics in terms of the query planner. For *PostgreSQL* it does not matter where the condition for joining two tables is placed. In other words, *PostgreSQL* does not distinguish between an inner, implicit or equi-join.

Besides, the queries are composed to join two TCP-H tables with different number of rows to investigate the join performance. The first two queries joined the `lineitem` table with the `orders` and the `part` table, respectivly. The last query joined the `orders` with the `customers` table.

[8] The SQL standard standardized an inner join as `<table a> [INNER] JOIN <table b> WHERE <a.xyz> = <b.xyz>`.

[9] An equi-join is in the form `<table a> JOIN <table b> ON <a.xyz> = <b.xyz>`. The SQL standard allows shorthand for the column to join by using the `USING` clause.

Finally we select the amount of joined rows by using the `count(*)` aggregate function. This forces *PostgreSQL* to use sequential scans for the mentioned tables.

As assumed, our test results indicate that the join performance is strongly related to the row scan performance. The test results are similar to the ones of our TPC-H and SSB benchmarks. This means they do not resemble much in their energy consumption but in their execution times. Unsurprisingly the lower the amount of rows to be scanned for joining, the lower is the execution time and therefore the energy efficiency is quite better. Although compared with query with the unconditional join mentioned above, the performance is fundamentally worse.

We were also interested in the effects of eleminating duplicates from a result set. For this purpose we formed a test query using the `DISTINCT()` clause on a column of the `lineitem` table not having a supporting index. The query plan revealed a sequence scan and the removal of duplicates by hashing the values. Again, the test results showed the same characteristics as all of our database tests performing a sequence scan.

4 Summary

At first, our tests with our database server using normal HDDs indicates an energy increase of roughly 9 W when the HDDs are fully utilized. Compared to the energy consumption of the other measured core components during the tests, this is insignificant. The argument, SSDs should be prefered because they consume up to 12 times less energy compared to HDDs, is invalid in this context.

As *Lang et all.* stated in the summary of [17], evaluating the energy efficiency of a DBMS needs the inclusion of entire workloads, not just single queries. This study makes use of three different and complete workloads that allows a more comprehensive look at the energy efficiency of a relational DBMS. Most of the benchmark queries caused a massive usage of sequential scans. This implies that the sequential read performance is an extremly important factor that affects the energy consumption. Actual SSDs clearly outperform normal HDDs but in this case enterprise grade HDDs can be used because they offer nearly the same performance as SSDs.

As mentioned in the introductionary section of this paper, there are more factors and not only technical parameters that influence the performance and thus the energy efficiency of a database server.

For example, the filesystem cache provided by the operating system is more relevant for the execution of a database query in *PostgreSQL* than its internal cache. Based on our experiments, we recommend not to assign more than 50 percent of the main memory to *PostgreSQL* for operations. With more assigned main memory the remaining processes of the operating system are forced to use the remaining portion. This causes the operating system to swap this portion to the hard drive which leads to a dramatic reduction of the performance.

Another important factor for the energy efficiency of the used database benchmark is the kind of accessing *PostgreSQL* (single vs. multiple database connections). Our assumption, subsequent queries of the benchmarks would benefit from *PostgreSQL*'s internal cache by using just a single database connection, does not come true. In fact, the opposite performed better.

Our tests also indicate the fact that energy saving settings are counterproductive for a database server that is reasonably utilized because it decreases the overall system performance.

References

1. Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D., Silberschatz, A., Rasin, A.: HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads. Proc. VLDB Endow. 2(1), 922–933 (2009), http://dl.acm.org/citation.cfm?id=1687627.1687731
2. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: A view of cloud computing. Commun. ACM 53(4), 50–58 (2010), http://doi.acm.org/10.1145/1721654.1721672
3. Barroso, L.A., Holzle, U.: The Case for Energy-Proportional Computing. Computer 40(12), 33–37 (2007)
4. Beckmann, A., Meyer, U., Sanders, P., Singler, J.: Energy-efficient sorting using solid state disks. In: Green Computing Conference 2010, pp. 191–202 (2010)
5. Graefe, G.: Database servers tailored to improve energy efficiency. In: Proceedings of the 2008 EDBT Workshop on Software Engineering for Tailor-Made Data Management, SETMDM 2008, pp. 24–28. ACM, New York (2008), http://doi.acm.org/10.1145/1385486.1385494
6. Härder, T., Hudlet, V., Ou, Y., Schall, D.: Energy efficiency is not enough, energy proportionality is needed! In: Xu, J., Yu, G., Zhou, S., Unland, R. (eds.) DASFAA Workshops 2011. LNCS, vol. 6637, pp. 226–239. Springer, Heidelberg (2011), http://dl.acm.org/citation.cfm?id=1996686.1996716
7. O'Neil, P.E., O'Neil, E.J., Chen, X.: The Star Schema Benchmark (SSB), revision 3 (2007), http://www.cs.umb.edu
8. Papadimitriou, C.H.: Computational complexity. In: Encyclopedia of Computer Science, pp. 260–265. John Wiley and Sons Ltd., Chichester, http://dl.acm.org/citation.cfm?id=1074100.1074233
9. Poess, M., Floyd, C.: New TPC benchmarks for decision support and web commerce. SIGMOD Rec. 29(4), 64–71 (2000), http://doi.acm.org/10.1145/369275.369291
10. Polte, M., Simsa, J., Gibson, G.: Comparing performance of solid state devices and mechanical disks. In: Petascale Data Storage Workshop, PDSW 2008, 3rd edn., pp. 1–7 (2008)
11. Schall, D., Hudlet, V., Härder, T.: Enhancing energy efficiency of database applications using SSDs. In: Proceedings of the Third C* Conference on Computer Science and Software Engineering, C3S2E 2010, pp. 1–9. ACM, New York (2010), http://doi.acm.org/10.1145/1822327.1822328
12. Schröder-Preikschat, W., Wilkes, J., Isaacs, R., Narayanan, D., Thereska, E., Donnelly, A., Elnikety, S., Rowstron, A.: Migrating server storage to SSDs. In: Proceedings of the 4th ACM European Conference on Computer Systems, EuroSys 2009, p. 145. ACM, New York (2009)

13. Harizopoulos, S., Shah, M.A., Meza, J., Ranganathan, P.: Energy Efficiency: The New Holy Grail of Data Management Systems Research. In: CIDR 2009 (2009)
14. Tsirogiannis, D., Harizopoulos, S., Shah, M.A.: Analyzing the energy efficiency of a database server. In: Proceedings of the 2010 International Conference on Management of data, SIGMOD 2010, pp. 231–242. ACM, New York (2010)
15. Wang, J., Feng, L., Xue, W., Song, Z.: A survey on energy-efficient data management. SIGMOD Rec. 40(2), 17–23 (2011), http://doi.acm.org/10.1145/2034863.2034867
16. Lang, W., Patel, J.M.: Towards Eco-friendly Database Management Systems. In: CIDR 2009 (2009)
17. Lang, W., Harizopoulos, S., Patel, J.M., Shah, M.A., Tsirogiannis, D.: Towards Energy-Efficient Database Cluster Design. CoRR abs/1208.1933 (2012)
18. Xu, Z.: Building a power-aware database management system. In: Proceedings of the Fourth SIGMOD PhD Workshop on Innovative Database Research, IDAR 2010, pp. 1–6. ACM, New York (2010), http://doi.acm.org/10.1145/1811136.1811137
19. Zheng, H., Zhu, Z.: Power and Performance Trade-Offs in Contemporary DRAM System Designs for Multicore Processors. IEEE Transactions on Computers 59(8), 1033–1046 (2010)
20. Xu, Z., Tu, Y.-C., Wang, X.: Exploring power-performance tradeoffs in database systems. In: ICDE 2010. pp. 485–496 (2010)

Chaining Data and Visualization Web Services for Decision Making in Information Systems

Ahmet Sayar[1,*] and Marlon E. Pierce[2]

[1] Kocaeli University, Computer Engineering Department
Umuttepe Campus, 41380, Kocaeli-Turkey
[2] Community Grids Laboratory, Indiana University
2719 East 10th Street Bloomington, IN, 47408, USA
ahmet.sayar@kocaeli.edu.tr, mpierce@cs.indiana.edu

Abstract. Decision making in information systems increasingly relies on analyses of data in visual formats which are created from distributed heterogeneous data belonging to the separate organizations. This paper presents distributed service architecture for creating and managing the production of knowledge from distributed collections of data sources through integrated data-views. Web Services provide key low level capability but do not define an information or data architecture. These are left to domain specific capabilities metadata and domain specific common data model. Datasets are considered to be defined with domain specific spatial and non-spatial attributes for displaying and querying. We propose blueprint architecture and define the principles and requirements for general information systems domains.

Keywords: Information Systems, Decision making, Web Services, Visualization, DIKW.

1 Introduction

The World Wide Web and its associated Web programming models have revolutionized accessibility to data/information sources. At the same time, numerous incompatible data formats, data heterogeneity (both the data types and storage formats), and machine un-readability of data limit data integration and federation [1]. The seamless integration and sharing of data from distributed heterogeneous data sources have been the major challenges of information system communities and decision support systems [2]. In order to be able to integrate and share data/information, data sources need to be in interoperable formats and provide standard service interfaces interacted with standard message formats and transport protocols. The interoperability issues have been studied by many public and private organizations over the last two decades at the data and service levels. Among these are the Web Service standards (WS-I) [3] for cross-language, platform and operating systems, and International Virtual Observatory Alliance (IVOA) [4] and Open

* Corresponding author.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 44–53, 2013.

Geospatial Consortium (OGC) [5] for defining domain specific data model and online service definitions in Astronomy and Geographic Information Systems (GIS), respectively.

Information systems are mostly used by decision makers coming from various kinds of domains and from various expert levels. Decision making increasingly relies on analyses of data in visual formats such as images, animations, etc. Most of the analyses include creation of integrated data views in which data sources provided by heterogeneous services. By this way decision makers query the heterogeneous data sources as a single resource. This removes the hassle of individually accessing each data source.

Decision making in Information systems mostly require integrating multiple heterogeneous data sets from various heterogeneous resources, and analyzing them from a single access point by visualization tools. There are many ad-hoc solution-frameworks developed for application specific purposes [6,7]. However, no one defines such a framework in terms of requirements and constraints. Katsis et al. [8] defines view-based integration issues from the point of global and local data format heterogeneities. They specify the relationship of sources with the global view. They determine that view-based integration systems follow global as view (GAV), Local As view (LAV) and Global and Local as View (GLAV) approaches. In GAV, the global database (schema) is expressed as a function of the local database (schema). LAV on the other hand follows the opposite direction. Finally GLAV is generalization of the two. Another closely related work is visual integration tool for heterogeneous data types by unified vectorization [9]. They propose a visual data integration technique to aid the end-users to identify semantically similar data attributes in a semi-automated fashion. This is achieved by using Self Organizing Map (SOM) to classify unfamiliar database entities. SOM is mostly used in text visualization related works, and applied in various projects.

We propose a framework for chaining data and visualization Web Services for decision making in information systems. The proposed framework is actually based on the idea of the Data-Information-Knowledge-Wisdom hierarchy (DIKW) [10,11]. In DIKW, wisdom is defined in terms of knowledge, knowledge in terms of information and information in terms of data. The system does not allow actual physical data integration. Physical integration results in both high data maintenance and storage costs. Therefore, it is better to keep data sets at their originating sources and maintain their internal structures. This enables large degree of autonomy. Individual data sets are integrated to the system through mediator services (see Fig. 1).

Web Services provide key low level capability but do not define an information or data architecture. These are left to domain specific capabilities metadata and domain specific common data model. Each-domain has different set of attributes for the data and its model is defined in core language. These are defined by standard bodies of the corresponding domain. Cross-domain interoperability issues are not handled in this paper.

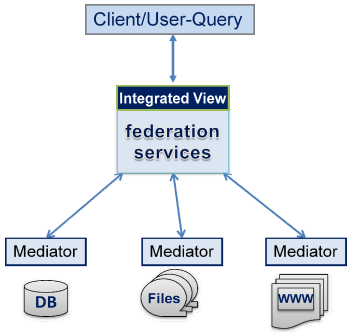


Fig. 1. Integrated data views

The remaining of the paper is structured as follows. Section 2 presents the proposed framework. Section 3 gives the details about the intercommunication and message exchange between the abstract components in the framework. Section 4 scrutinizes the three different information systems domains (GIS, Astronomy and Chemistry) and shows to what extent the proposed framework can be realized. Section 5 presents a sample ASIS framework for GIS domain, and Section 6 concludes the paper.

2 Application Specific Information Systems (ASIS)

This paper proposes blueprint architecture in terms of principles and requirements and calls it Application Specific Information System (ASIS). Developing such a framework requires first defining a core language expressing the primitives of the domain; second, key service components, service interfaces and message formats defining services interactions; and third, the capability file requirements (based on core-language) enabling inter-service communications to chain (link) the services for the federation (see Fig. 2).

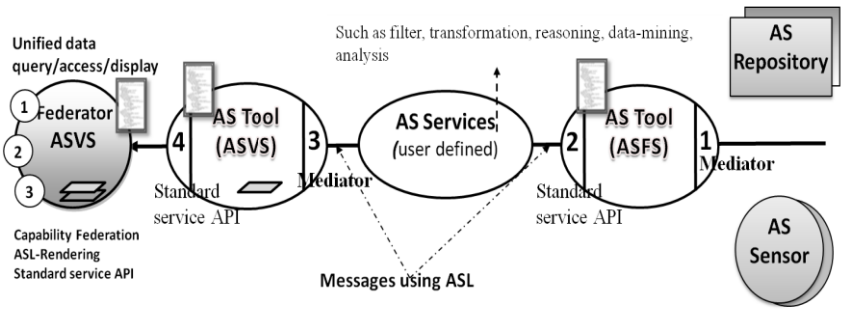


Fig. 2. Application Specific Information System (ASIS)

The architecture proposed in this paper is a high level architecture that consists of abstract components and explains their data flow and components' interactions. In this section, we focus on the principles and requirements in terms of data and service interfaces in any information system domains. It should be noted that this abstract architecture is intended to be domain-specific. That is, it may be realized in chemistry or astronomy, for example, but we are not suggesting cross-domain interoperability.

ASIS consists of two groups of filter-like Web Services. These are Application Specific Feature Service (ASFS), which is actually a data service, and Application Specific Visualization Service (ASVS), which is actually a display service. These services also behave as mediators. ASFS perform their mediation services with adaptors converting any data to common data model encoded in application specific language (ASL) (number 1 in Fig. 2). ASFS also mediate resource specific heterogeneous service interfaces with standard service interfaces (number 2 in Fig. 2). Standard service API enables querying on data (e.g. GetASL) and their metadata (e.g. GetCapabilities). On the other hand, ASVS mediate any data in common data model (ASL) and convert into a visual format (an image) with pre-defined attributes such as bounding boxes, projections and resolutions (number 3 in Fig. 2). ASVS also enables querying of ASL through visual image representations through standard service interfaces (number 4 in Fig. 2). Those interfaces enable performing attribute-based querying of ASL through displays. ASVS also provide service interfaces to list the available ASL data with their attributes, which enables clients to make appropriate queries.

ASIS is a proposed solution to heterogeneous data integration. This solution enables inter-service communication through well-defined service interfaces, message formats and capabilities metadata. Data and service integration is done through capability federation of these services. ASFS and ASVS are Web Services, and each service is described by corresponding generic metadata descriptions that can be queried through Web Service invocations. In addition to allowing service discovery, this approach also enables at least three important qualities of services. First, services of the same type that provide a subset of the request can be combined into a "super-service" that spans the query space and has the aggregate functionality of its member services. Second, the capability metadata can be used to determine how to combine services into filter chains with interconnected input-output ports. Third (and building on the previous two), capabilities of super-services can be broken into smaller, self-contained capabilities that can be associated with specific services. This enables performance gains through load-balancing.

ASFS and ASVS are filter-like Web Services and communicate with each other through capability metadata exchange interface. Capability metadata defines service and data attributes, and their constraints and limitations to enable clients to make valid queries and get expected results. Being a Web Service enables filter services to publish their interfaces, locate each other and chain together easily. Filters have inter-service capabilities and are chainable. If the filter is capable of communicating and obtaining data from other filters, and updates (or aggregates) its capability metadata

with these data (after capability files exchange), then it can claim that it serves these data. Filter Services are information/data services that enable distributed data/information access, querying and transformation through their predictable input/output interfaces defined by capability document. Filter located in the same community network can update their capability metadata dynamically through the standard service interface of the filters (e.g. getCapabilities). In addition, dynamically updating capabilities of filters enable removal of obsolete data or down filters. Capabilities metadata and ASL are closely related to each other. One defines the domain-specific data and other defines the query and response constraints over the service and data provided.

ASVS must visualize information and provide a way of navigating ASFS and their underlying database. ASVS must provide human readable information such as text and graphs (scalable vector graphic (SVG) [12] or portable network graphic (PNG) [13]) images. These are final outputs of the system and might be called as knowledge. ASVS defines the knowledge (formats, types etc.) in terms of what data needs to be rendered from what services. Services are basically ASFS and possibly other ASVS. This definition is done in ASVS's capability metadata. An ASFS is an annotation service providing heterogeneous data in common data model with an attribute-based query capability. ASFS serves data in ASL, which must be realized as a domain specific XML-encoded common data model containing content and representation tags. Heterogeneity in queries and data formats is handled through resource specific mediators. ASL is a domain specific encoding of common data defining the query and response constraints over the service and data provided.

A sample scenario to apply the presented framework can be given in GIS domain. Pattern Informatics (PI) [14] is a project on earthquake forecasting. In PI, data sets are earthquake seismic data records collected from sensors. ASFS provide data records in global data format with attribute and spatial based standard queries. ASFS overlay seismic data records and earthquake forecast values [15] on satellite map images and enable interactive visual querying. Depending on the magnitudes of expected earthquake seismicity, maps are also possibly overlaid with hot-spot plots in colored boxes.

3 Inter-service Communications

Inter-service communication is achieved through common service interfaces and capability metadata exchange. The standard service interfaces can be grouped into three types: a) capability metadata exchange: inter-service communication (set-up stage); b) interactive data display: selecting layer composition and bounding box regions; and c) querying of data itself over the display, getting further information about the data content and attributes.

As mentioned before, capability helps clients make valid requests for its successive queries. Capability basically provides information about the data sets and operations available on them with communication protocols, return types, attribute based constraints, etc. Each domain has different set of attributes for the data and it is defined in ASL common data model. For example, in GIS domain, attributes might be

bounding box values (defining a range query for data sets falling in a rectangular region) and coordinate reference system.

Standard requests/query instances for the standard service interfaces are created according to the standard agreed-on request schemas. These are defined by open standards bodies in corresponding domains. The request instances contain format and attribute constraints related to the ASL common data model. For example in the GIS domain, getMap request defines a map images' return format (JPEG, PNG, SVG, etc.), height, width, bounding box values, and so on. Format, height and width are related to display, but bounding box values are related to the attributes of the data defined in its ASL representation provided by ASFS. In this specific example of the getMap request, ASVS must both visualize information through the getMap service interface and provide a way of navigating ASFS services and their underlying database. ASVS make successive queries to the related ASVS to get the ASL data and render it to create final display for its clients.

In ASIS, the task of mediators is to translate requests to the standard service interfaces to those of the information/data sources', and transform the results provided by the information source back to the ASIS's standard formats. For ASFS, the returned data is ASL, and for ASVS the returned results can be any kind of display format such as images.

Acting as a proxy of information source, the mediators communicate with an information source in its native language and API. They communicate with ASIS in a commonly agreed language (ASL) and Web Service API calls. Because of the obvious heterogeneity between different science domains, each will need to extend and create its own service interfaces, as well as queries for those services. Using common data model-ASL and common services make autonomous resource to be integrated into the system in a manageable way.

The mediators-wrappers enable data sources integrated to the system conform to the global data model (ASL) but enable the data sources to maintain their internal structure. At the end, this whole mediator system provides a large degree of autonomy. Instead of actual physical data federation, system makes distributed querying and response composition on the fly. Mediators perform query and data format conversions.

4 Abstract Components and Matching to Sample Science Domains

GIS is a mature domain in terms of information system studies and experiences. It has standards bodies defining interoperable online service interfaces and data models such as OGC (Open Geospatial Consortium) and ISO/TC211, but many other fields do not have this. We have surveyed two science domains (Astronomy and Chemistry). Table 1 presents the results briefly in terms of service counterparts (ASIS vs. science domains).

Astronomy has a standards body, the International Virtual Observatory Alliance (IVOA), for defining data formats and online services that are somewhat analogous to

the OGC standards. FITS (Flexible Image Transfer), Images and VOTable [16] are the data models. SkyNodes are database servers with an ADQL (Astronomy Distributed Query Language) based SOAP interfaces that return VOTable-encoded results. VOPlot and TopCat are two services to visualize the astronomy data in the format of VOTable, FITS and images. VOResource and UCD are the metadata definition and standards for the service descriptions [17].

Chemistry, although a vastly different field, does provide a common data model (CML[18]) that can be used to build up Web Services. Although many research groups have investigated service architectures for chemistry and chemical informatics, the field has (to our knowledge) no Web Service standards-defining body equivalent to the OGC or IVOA.

Table 1. Components and common data models matching for sample science domains

	ASIS				
Science Domains	**Common Data Model (ASL)**	**Components ASFS**	**ASVS**	**Metadata**	**Standard Bodies**
GIS	GML	WFS	WMS	capability.xml schema	OGC and ISO/TC211
Astronomy	VOTable, FITS	SkyNode	VOPlot TopCat	VOResource	IVOA
Chemistry	CML, PubChem	None	NO standard JChemPaint, JMOL	None	None

5 ASIS Application to GIS Domain

Data in GIS are stored in various places using various formats. These formats may be closed (i.e., proprietary dependent) or opened, normalized or not. A consequence is that the analysis and/or decision-making may be difficult or time-consuming due to this heterogeneity of formats. To avoid these drawbacks and in order to facilitate accesses and uses of these data, the Open Geospatial Consortium (OGC) defined some propositions. These propositions allow accesses to geographical data without taking into account physical aspects of communications (i.e., with an URL). They also propose a set of parameters, methods and communication rules to simplify accesses and manipulations as soon as clients and servers respect them. The Web Map Service (WMS) and Web Feature Service (WFS) are two well-known web services defined by OGC standards. WMS deal with dynamic production of maps as images, built with geo-referenced data. WFS are designed to provide an access and a manipulation tool of geographical data within a map. WMS and WFS provide standard service interfaces and their operations provide data associated with a geographical map or object. Information may be alphanumerical such as a town name or graphical such as the segments defining a border of a town.

WMS correspond to ASVS in the proposed framework. WMS standard induces three main operations. Fig. 3 presents these operations: GetCapabilities, GetMap, GetFeatureInfo. The most important facility for WMS is to provide a map from its available layers. Parameters such as desired layers, representation styles, size, relevant geographical areas, the projection system or the output format must be provided to the GetMap operation. The obtained result is a map, which is an image that can be displayed by conventional web browsers. This operation corresponds to a visualization process.

The GetFeatureInfo operation corresponds to a query process. An end-user does not manipulate structured data but an image. To be able to access data, an end-user must select an object on the image. The GetFeatureInfo operation provides this selection. Nevertheless, some limits appear in the sense that the provided schema is the set (or a sub-set) of available data for this/these object(s). No link is performed with the environment of this object. Furthermore, an end-user must be aware of the database schema in order to select only a sub-set of available data. This information is kept in Capabilities.xml document and provided by GetCapabilities service interface upon requests.

WFS in Fig. 3 correspond to ASFS in the proposed framework. WFS are OGC specifications for data access. It describes responses of a Web server to geographical data manipulation operations. These operations are based on the CRUD manipulations of geographic data based on alphanumeric/spatial constraints: creation (C), read (R), update (U) and deletion (D). The formalism used to model data exchanges for the WFS specification is GML (Geography Markup Language). GML correspond to ASL in the proposed framework. GML is a common data model for geographic data, and defined by OGC specifications. It is an XML dialect and is designed to encode, to manipulate and to exchange geographical data. The WFS specification defines five operations to send queries to a geographical data server and to get answers from it: GetCapabilities, DescribeFeatureType, GetFeature, Transaction and LockFeature. Fig. 3 presents the basic operations of the WFS specification.

The most important operation for a WFS server is the GetFeature operation. This operation delivers data instances typed by features, identifies properties that should be delivered and provides the results of spatial and non-spatial queries. Selection criteria are defined by using a filter. To get relevant data, a client specifies an object identifier in the filter. The WFS server receives the GetFeature query, determines the correct database, creates and sends a SQL statement to the database and formats the results. The filter is used to manage the “Where” clause of a SQL statement. Spatial data are handled with the spatial schema defined in the ISO 19107 norm. Before sending results to a client, the server transforms objects using a GML format.

In summary, Fig. 3 illustrates an application of ASIS in GIS domain. Such a framework has been used in some Geo-science projects [19-21] with application specific extensions and modifications.

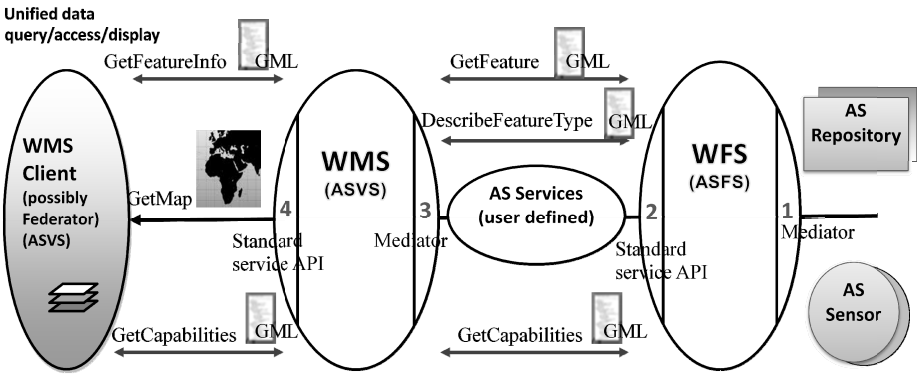


Fig. 3. Application of ASIS to GIS Domain

6 Conclusion

We have presented a service-oriented architecture for understanding and managing the production of knowledge from the distributed observation, simulation and analysis data through integrated data-views in the form of images. It enables unified data access/query and a display from a single access point. The infrastructure is based on a common data model, standard Web-Service components for data and visualization, and a federation service. The framework applies just-in-time (late-binding) federation in which the data is kept in its originating sources all the time. This enables autonomy and easy data maintenance.

The architecture proposed in this paper is a blueprint architecture defining the principles and requirements for application specific information systems for general science domains. We focus on the principles and requirements in terms of data and service interfaces in any information system domains. We have defined two general service types (ASFS and ASVS) with limited number of service interfaces. Service interfaces enable metadata exchange and data querying. Data flows from databases to users through ASFS and then ASVS. It should be noted that this abstract architecture is intended to be domain-specific. That is, it may be realized in chemistry or astronomy, for example, but we are not suggesting cross-domain interoperability. Due to the domain specific data heterogeneity, each domain should define its own ASL and corresponding queries.

References

1. Huang, Y., Cai, G., Wang, G.: Dynamic Service Composition Based on Federated Agents. Energy Procedia 13, 5536–5543 (2011)
2. Sayar, A., Pierce, M., Fox, G.C.: Unified Data Access/Query over Integrated Data-views for Decision Making in Geographic Information Systems. In: Bessis, N. (ed.) Grid Technology for Maximizing Collaborative Decision Management and Support: Advancing Effective Virtual Organizations, p. 360. IGI Global - Information Science Reference Bedfordshire, UK (2009)
3. WS-I: WS-I (2012), `http://www.ws-i.org/` (accessed March 23, 2012)

4. IVOA: International Virtual Observatory Alliance (2012), http://www.ivoa.net/ (accessed January 12, 2013)
5. OGC Schema (2008), http://schemas.opengis.net/ (accessed September 14, 2008)
6. Chan, B., Talbot, J., Wu, L., Sakunkoo, N., Cammarano, M.: Vispedia: Ondemand Data Integration for Interactive Visualization and Exploration. In: ACM SIGMOD International Conference on Management of Data, Rhode Island, USA, une 29-July 2, pp. 1139–1142. ACM (2009)
7. Cammarano, M., Dong, X.L., Chan, B., Klingner, J., Talbot, J., Halevy, A., Hanrahan, P.: Visualization of Heterogeneous Data. IEEE Transactions on Visualization and Computer Graphics 14(6), 1213–1220 (2008)
8. Katsis, Y., Papakonstantinou, Y.: View-based Data Integration. Encyclopedia of Database Systems, 3332–3339 (2009)
9. Bourennani, F., Pu, K.Q., Zhu, Y.: Visual Integration Tool for Heterogeneous Data Type by Unied Vectorization. In: IEEE International Conference on Information Reuse & Integration (IRI 2009), Las Vegas, NV, pp. 132–137. IEEE (2009)
10. Zins, C.: Conceptual approaches for defining data, information, and knowledge. Journal of the American Society for Information Science and Technology 58(1), 479–493 (2007)
11. Rowley, J.: The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science 33(2), 163–180 (2007), doi: 10.1177/ 0165551506070706
12. Andersson, O., et al.: Scalable Vector Graphics (SVG) Specification Version 1.1. In. World Wide Web Consortium, W3C (2003)
13. Adler, M., Boutell, T., Bowler, J., Brunschen, C., Costello, A.M., Crocker, L.D., Dilger, A., Fromme, O.: Gailly, J.-l., Herborth, C.: Portable Network Graphics Specification (PNG). REC-PNG-20031110 (2003)
14. Holliday, J.R.: Chen, C.-C., Tiampo, K.F., Rundle, J.B., Turcotte, D.L., Donnellan, A.: A RELM Earthquake Forecast Based on Pattern Informatics. Paper Presented at the American Geophysical Union (AGU) - fall meeting, San Francisco, California (December 2005)
15. Rundle, J.B., Turcotte, D.L., Shcherbakov, R., Klein, W., Sammis, C.: Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. Geophysics 41(4) (2003), doi:10.1029/2003RG000135
16. Williams, R., Ochsenbein, F., Davenhall, C., Durand, D., Fernique, P., Giaretta, D., Hanisch, R., McGlynn, T., Szalay, A., Wicenec, A.: VOTable: A Proposed XML Format for Astronomical Tables. In. US National Virtual Observatory (2002)
17. Yasuda, N., Mizumoto, Y., Ohishi, M., O'Mullane, W., Budavári, T.A., Haridas, V., Nolan Li, T., Malik, A.S., Hill, M., Linde, T., Mann, B., Page, C.: Astronomical Data Query Language: Simple Query Protocol for the Virtual Observatory. Paper presented at the Astronomical Data Analysis Software and Systems XIII. ASP Conference Series, San Francisco, USA
18. Holliday, G.L., Murray-Rust, P., Rzepa, H.S.: Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions. Journal of Chemical Information and Modeling 46, 145–157 (2006)
19. Aydin, G., Sayar, A., Gadgil, H., Aktas, M.S., Fox, G.C., Ko, S., Bulut, H., Pierce, M.E.: Building and Applying Geographical Information Systems Grids. Concurrency and Computation: Practice and Experience 20(14), 1653–1695 (2008)
20. Sayar, A.: High Performance Federated Service Oriented Geographic Information Systems. PhD (2009), http://gradworks.umi.com/33/44/3344771.html
21. Pierce, M.E., Fox, G.C., Aktas, M.S., Aydin, G., Qi, Z., Sayar, A.: The QuakeSim Project: Web Services for Managing Geophysical Data and Applications. Pure and Applied Geophysics (PAGEOPH) 165(3-4), 635–651 (2008), doi:10.1007/s00024-008-0319-7

A Fully Reversible Data Transform Technique Enhancing Data Compression of SMILES Data

Shagufta Scanlon and Mick Ridley

University of Bradford, Bradford, UK
s.a.scanlon@student.bradford.ac.uk, m.j.ridley@bradford.ac.uk

Abstract. The requirement to efficiently store and process SMILES data used in Chemoinformatics creates a demand for efficient techniques to compress this data. General-purpose transforms and compressors are available to transform and compress this type of data to a certain extent, however, these techniques are not specific to SMILES data. We develop a transform specific to SMILES data that can be used alongside other general-purpose compressors as a pre-processor and post-processor to improve the compression of SMILES data. We test our transform with six other general-purpose compressors and also compare our results with another transform on our SMILES data corpus, we also compare our results with untransformed data.

Keywords: SMILES, Data Transform, Data Compression.

1 Introduction

The Simplified Molecular Input Line Entry System (SMILES) language was developed to represent two-dimensional molecular structures in a concise and compact way allowing for storage and processing improvements. General-purpose compressors allow for further reductions in storage and processing costs [23], [8].

With the continuous expansion of chemical databases [15] and the need for efficient storage and searching of molecular structure representations, such as SMILES [23], [8], storage and processing costs of these representations need to be further improved.

Data can be transformed by exploiting the specific information contained in the data to its advantage. Transformed data can be used alongside general-purpose compression techniques to further improve compression results [21], [4], [20].

We make the following contributions in this paper:

- We present our SMILES-specific transform designed to enhance the compression of SMILES data when used with other general-purpose compressors.
- We provide results from using general-purpose compression techniques on a breakdown of different SMILES transform scenarios and a combination of techniques used in our SMILES transforms.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 54–68, 2013.

- We provide a comparison of the results from our SMILES transforms with Word Replacing Transforms (WRT) [21], [12], [20] optimized for different compression algorithms and also on untransformed data.

The remainder of this paper is structured as follows. Section 2 discusses molecular structure representations and how compression has been used on SMILES data. Sections 3 to 5 introduce the SMILES chemical language, general-purpose compression and data transformation techniques, respectively. Section 6 illustrates our SMILES-specific transform design, grammar and architecture. Section 7 provides details of the experiments conducted, including the data collected, testing environment, metrics computed, experiment methodology and results. The results shown demonstrate the benefits of using compression over transformed data. The actual transformation results themselves have been omitted to help maintain the focus of this paper, which is to improve the compression of SMILES data using SMILES-specific transforms, and also due to space limitations. Section 8 discusses the key findings and concludes this paper.

2 Molecular Structure Representations

Several molecular structure representations have been implemented over the years. The most popular linear representations include SMILES, which is human-readable, and International Chemical Identifier (InChI), which is machine-readable. Other linear representations include Wiswesser Line Notation (WLN), Representation of Organic Structures Description Arranged Linearly (ROSDAL), SYBYL Line Notation (SLN), Modular Chemical Descriptor Language (MCDL) and InChIKey [14], [18], [10]. Other types of representations include MOL and SDF formats for single and multiple molecules, respectively. MOL and SDF consist of coordinates and connections between atoms [14]. The authors in [8] suggested that SMILES representations use between 50% and 70% less storage space in comparison to their equivalent connection tables [8].

The authors in [14] developed a molecular structure representation using barcodes in SMILES format. Barcodes were chosen to handle errors and SMILES was chosen due to its human-readability and compact nature in comparison to other molecular representations, such as MOL and SDF. However, they argued that in the future with the potential increase in error correction levels required, the compression of SMILES would be inevitable to avoid an increase in data storage. They discussed the use of Lempel-Ziv-Welch (LZW), a lossless dictionary-based compression algorithm, and found that data reduction with LZW can be achieved approximately by a factor of two. As well as compression, the authors also discussed the use of a more compact representation to SMILES instead, known as the Automatic Chemical Structure (ACS). ACS provides a condensed representation of a molecular structure by assigning unique identifiers to molecular substructures or fragments. However, as molecular information from a SMILES representation is omitted from the ACS encoded barcode, backward mapping from ACS to SMILES is important [14].

3 SMILES Data

SMILES is a linear chemical notation language originally developed by Weininger in 1988 [23] and further extended by Daylight Chemical Information Systems [8]. It is used by practitioners in the Chemoinformatics field to represent a linear form of two-dimensional molecular structures [23].

Table 1 provides a summary of the basic representation rules for SMILES notations and some examples. SMILES notations are essentially represented using ASCII characters in linear format with no spaces. Molecular structures can have several different and valid SMILES representations and SMILES representations are not unique. Although, unique representations can be generated using a canonicalization algorithm [23], [8].

Atomic symbols are used to represent atoms. Non-organic atoms can be represented inside square brackets, also included in the brackets are the number of Hydrogen atoms and charges declared for the atoms. The range of elements within the organic subset can be depicted with the square brackets omitted, this indicates that Hydrogen atoms are present for these atoms even without the square brackets. Aliphatic and aromatic atoms are represented in uppercase and lowercase characters, respectively [23], [8].

Neighboring atoms are attached to each other by single bonds or aromatic bonds, which can be excluded, unless otherwise specified by double or triple bonds. Branches are parenthesized and can be characterized as nested or stacked. To symbolize a cyclic structure, a single bond must be broken inside a cyclic ring. Cyclic ring opening and closing are determined by numbers assigned following the ring opening and ring closing atomic symbols. A period characterizes the separation of disconnected structures [23], [8].

Table 1. Generic SMILES Representation Rules and Examples [23], [8]

Generic SMILES Rules	Example Representations
Non-Organic Atoms	[S], [H+]
Aliphatic Organic Atoms	B, C, N, O, P, S, F, Cl, Br, I
Aromatic Organic Atoms	b, c, n, o, p, s
Single Bonds	C-C, CC
Double Bonds	C=C
Triple Bonds	C#N
Aromatic Bonds	c:c, cc
Nested or Stacked Branches	C=CC(CCC)C(C(C)C)CCC
Ring Closures	C1CCCCC1
Disconnections	[Na+].[0-]c1ccccc1

4 General-Purpose Data Compressors

General-purpose compressors can be used to universally compress data of textual format [4], [20]. Table 2 shows a summary of the compression algorithms and techniques used in each general-purpose compressor used in our experiments.

Lossless compression techniques preserve the integrity of data by ensuring the original data is fully reconstructed from the compressed data on decompression [20]. Two common forms of lossless compression techniques are statistical and dictionary-based approaches. Statistical approaches to compression include the Huffman encoding algorithm which is a technique that involves the substitution of frequent characters with shorter codewords [4], [20]. Dictionary-based approaches involve the substitution of words or phrases with their corresponding indices in the dictionary. The Lempel-Ziv compression schemes which include LZ77, LZ78, LZW and LZMA to name a few, are all dictionary-based approaches [4], [19], [20].

Within dictionary-based approaches, static dictionaries can be suitable when information about the data being processed is already known and can be prepared in advance. They have the potential to improve compression results as the technique is one-pass, so an information gathering phase prior to compression would not be required, and the information would be readily available in the dictionary to process. However, the fixed nature of the dictionaries have the potential to negatively impact storage costs. A semi-static dictionary approach uses two-passes, the first to gather statistics from the data source and prepare the dictionary and the second to actually compress the data. The adaptive dictionary technique dynamically rebuilds the dictionary during compression [4], [20].

Table 2. General-Purpose Compressors

Compressors	Compression Algorithms/Techniques
7Zip [1], [19]	Back-End: LZMA (Default), LZMA2, PPMd, BZip2, DEFLATE
BZip2 [19]	Uses Burrows-Wheeler Block Transformation (BWT) and Huffman
GZip [19]	Based on DEFLATE (LZ77 and Huffman)
PPMd [22], [7], [20]	Based on Prediction by Partial Matching (PPM), Adaptive Statistical Technique using Context Modelling and Prediction
PPMVC [22], [12], [20]	Based on PPM with Variable-Length Contexts
ZPAQ [24], [17]	An Extension to PPM, Back-End: LZ77 (Default), BWT

5 Data Transform Techniques

Data transformation can be added as an extra pre-processing and post-processing step during data compression to enhance compression results. Where general-purpose compressors can be used to treat all types of data as text and compress them accordingly, modelling a data transformation technique on a specific type of data can provide far better compression results when used alongside other compressors [21], [4], [20].

Data-specific transforms have been developed for this purpose for different types of data. For example, to transform textual data, WRT was developed by Skibiński [21], [12], [20] to pre-process English language text based on matching words in the dictionary and replacing them with codewords. It uses capital conversion techniques, dictionary sorting according to word frequency, q-gram replacement whereby frequent q consecutive characters are replaced with shorter symbols, End-of-Line (EOL) symbol conversion to spaces, data protection techniques such as using a data filter, and encloses words with spaces. WRT was extended to Two-Level Word Replacing Transform (TWRT) to add data dictionaries that were specific to the type of data being processed, for example a dictionary containing commands from a Java programming language [21], [20]. WRT transforms optimized for BWT, LZ77, PAQ and PPM [12], [20] have been compared with our SMILES transforms in our experiments.

6 SMILES-Specific Data Transform

We have developed a transform language specific to SMILES notations in order to improve the compression of SMILES data. The design and techniques used in this transform was aimed at providing a good balance between data storage and processing costs, whilst maintaining data integrity. As with the use of general techniques such as the substitution of n-grams and the addition of prefixes, SMILES-specific techniques such as substituting atomic symbols with their equivalent atomic numbers was used. The atomic symbol conversion to atomic numbers, used in stage three of our transform, has an advantage as it allows the practitioner to process the data in its transformed state on decompression.

6.1 SMILES Transform Properties

The following highlights the properties held by the developed transform:

- Storage Costs Reduction – Ensures a reduction in storage costs when the transformed data is compressed with other general-purpose compressors.
- Processing Time Reduction – Compression and decompression times are reduced when compression is used over the transformed data.
- No Ambiguity – Handles any ambiguous SMILES tokens, particularly with aromatic elements.
- Fully Reversible – Data is fully reversible with data integrity preserved.

6.2 SMILES Transform Phases

The following shows the techniques used in the different phases of the transform:

1. N-Grams[1] – Frequent SMILES tokens from two to eleven characters in length are replaced with either other frequently used characters already used in the data, or with other unused characters.

[1] The concept of using N-Grams was taken from [5], [20] and [21].

2. Number Prefixes – Prefixes are added to SMILES tokens of numerical format using either other frequently used characters in the data or other unused characters. Note that this step is carried out prior to the next phase in order to avoid ambiguity and conflicts with atomic numbers.
3. Atomic Numbers and Prefixes – SMILES tokens that contain atomic symbols are matched and converted to their corresponding atomic numbers in the periodic table to ensure that there is no ambiguity. Atomic numbers are also prefixed with either other frequently used characters in the data ensuring no ambiguity or other unused characters. Aromatic elements are also converted to their corresponding atomic numbers to ensure consistency and are transformed last preserving the no ambiguity property.

6.3 SMILES Transform Representations

Table 3 illustrates the grammar used for our SMILES transform. Note that 2-grams have only been used for testing the first transformed representation to reduce entropy. In order to distinguish between aromatic elements and other elements, it is assumed that SMILES tokens that contain any aromatic elements would only do so in a manner that they do not conflict and cause direct ambiguity with any other elements. For example, the atomic symbol for Cobalt is Co, if this symbol was to appear in a SMILES string, then for our transform it has been assumed that this would be a two-letter symbol for Cobalt and not symbols for aliphatic Carbon, C, and aromatic Oxygen, o.

Table 3. SMILES Transform Grammar Applied to Different Scenarios[2]

Transform Scenarios	SMILES Tokens	1st Transform Grammar	2nd Transform Grammar	3rd Transform Grammar
2-Grams	CC	¬		
3 ... 11-Grams	CCC ... CCCCCCCCCCC	¬ ... ¬	() ... ()	
2-Grams	=C	|		
3 ... 11-Grams	C=C ... C1=CC=CC=C1	| ... |	[] ... []	
Number Prefixes (n)	n	£n	==n	
Element (e) Conversion and Atomic Number (n) Prefixes	e	;n, ~n, :n, <n, >n	*n, **n, ***n, ****n, *****n	[[n, [=[n, [=]n	[[[n, (=)n,

[2] The prefixes added to the atomic numbers in the second transform grammar was based on the star encoding '*' scheme proposed in [16] and also described in [20].

The following are examples of SMILES data and their equivalent transform representations to illustrate the transform scenarios further:

- 7-Grams:
  - SMILES: C1CC(CNCCCCCCC)CCC1CNCCCCCCC
  - 1st Transform: C1CC(CN¬)CCC1CN¬
  - 2nd Transform: C1CC(CN())CCC1CN()
  - SMILES: C(N)(=O)OC(C#C)(C1=CC=CC=C1)C2=CC=C(Cl)C=C2
  - 1st Transform: C(N)(=O)OC(C#C)(lC=C1)C2=CC=C(Cl)C=C2
  - 2nd Transform: C(N)(=O)OC(C#C)([]C=C1)C2=CC=C(Cl)C=C2
- Number Prefixes:
  - SMILES: C1=CC=CC(=C1)CCN(C)N=O
  - 1st Transform: C£1=CC=CC(=C£1)CCN(C)N=O
  - 2nd Transform: C==1=CC=CC(=C==1)CCN(C)N=O
- Atomic Numbers:
  - SMILES: CCC(Br)(CC)C(=O)NC(=O)NC(C)=O
  - 1st Transform: ~6~6~6(:35)(~6~6)~6(=~8)~7~6(=~8)~7~6(~6)=~8
  - 2nd Transform: *6*6*6(**35)(*6*6)*6(=*8)*7*6(=*8)*7*6(*6)=*8
  - 3rd Transform: [[6[[6[[6([[[35)([[6[[6)[[6(=[[8)[[7[[6(=[[8)[[7[[6([[6)=[[8

6.4 SMILES Transform Architecture

Figure 1 illustrates the SMILES transform architecture developed in Java. The transform parses the input SMILES tokens, stores the converted n-grams into a dictionary, uses regular expressions for adding prefixes to numbers, and stores the converted atomic elements along with their prefixes in dictionaries. The dictionaries created are static as the information is already known to us regarding which n-grams to use and the atomic numbers to use to replace the atomic symbols. It is intended that this technique will be extended to use semi-static or adaptive dictionaries in future work to further improve compression.

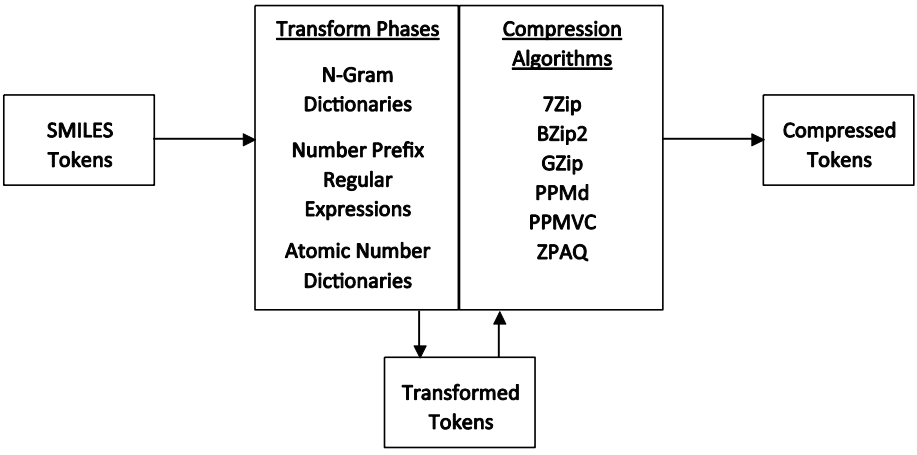


Fig. 1. SMILES Transform Architecture

7 Experiments

7.1 Data Collection

The experiments in this study were carried out on SMILES data extracted from 41 publicly available toxicology datasets. The datasets[3] were obtained from Bergstrom Melting Point [6], Carcinogenic Potency Database [6], CPDB [13], DSSTox[4] [9], Fathead Minnow Acute Toxicity [6], FDA's Carcinogenicity Studies [6], Fontaine Factor Xa [6], HERG [6], Huuskonen [6], Karthikeyan Melting Point [6], Li Blood-Brain-Barrier Penetration [6], National Toxicology Program [6], Stahl [6] and ToxBenchmark [2]. The SMILES data from all these files were combined into a single file and then multiplied to create a total of 12 files of sizes ranging from 1.71MB to 20.48MB. These files formed the basis of the data used in our experiments. In total, our data corpus consisted of 384 files. 288 files were used to provide a breakdown of our SMILES transform scenarios, these included 228 n-gram, 24 number prefix and 36 periodic number transformed files. 96 files were used to compare our SMILES transforms with other transformed and untransformed files, these included 36 SMILES and 48 WRT transformed files and 12 untransformed files.

7.2 Testing Environment

The following testing environment was used for the experiments:

- Operating System: Windows 7 Professional 64-bit OS.
- Processor: Intel(R) Core(TM) i5-3320M CPU @ 2.60GHz.
- Memory: 6.00GB (5.88GB usable).

7.3 Compression Metrics

The following metrics were computed in Java for these experiments:

- Compression Ratios – The size of the compressed files divided by the size of the uncompressed files.
- Compression Times – The average time taken for each compressor to compress the transformed files and untransformed files.
- Decompression Times – The average time taken for each compressor to decompress the compressed files.

7.4 Methodology

The compressors used on all transformed files were 7Zip [1], BZip2 [3], GZip [11], PPMd [7], PPMVC [12] and ZPAQ [24]. All compressors were used with their

[3] All whitespaces were removed.

[4] SMILES data was taken from the Structure SMILES column rather than the Parent SMILES column. Any incomplete or non-SMILES strings were removed.

default settings, it is intended that these settings will be changed in future work to provide for more detailed storage and processing results and further comparisons. The following SMILES-specific transforms were tested for all the compression metrics:

- 2-Gram_Chars to 11-Gram_Chars – N-grams using unused characters.
- 3-Gram_Brackets to 11-Gram_Brackets – N-grams reusing existing characters.
- Prefix_Chars – Number prefix addition using unused characters.
- Prefix_Equals – Number prefix addition reusing existing characters.
- Periodic_Chars – Atomic symbol to number conversion using unused characters.
- Periodic_Stars – Atomic symbol to number conversion using unused characters.
- Periodic_Brackets – Atomic symbol to number conversion reusing existing characters.
- BestStorage – This transform provides the best transformed file size and uses 2-gram_chars, prefix_chars and periodic_chars in its transform.
- AvgStorage – This transform provides an average transformed file size and uses 8-gram_chars, prefix_chars and periodic_stars in its transform.
- WorstStorage – This transform provides the worst transformed file size and uses 10-gram_brackets, prefix_equals and periodic_brackets in its transform.

Compression metrics were also computed for the following to compare our SMILES-specific transforms against:

- WRT-BWT [12] – WRT transform optimized for BWT compression.
- WRT-LZ77 [12] – WRT transform optimized for LZ77 compression.
- WRT-PAQ [12] – WRT transform optimized for PAQ compression.
- WRT-PPM [12] – WRT transform optimized for PPM compression.
- Untransformed – Untransformed data.

7.5 Results

Figures 2 to 10 illustrate the results and the key findings have been highlighted below:

- SMILES Transform Scenarios – The periodic_brackets transform scenario provided the best compression ratios for all compression algorithms except for PPMVC, where prefix_equals provided the best compression ratios for this algorithm. 2-gram_chars provided the worst compression ratios for all compression algorithms except for PPMVC, where 4-gram_chars gave the worst compression ratios for this algorithm. 2-gram_chars provided the best compression times for 7Zip, GZip, PPMVC, ZPAQ, along with 3-gram_chars for BZip2, and 8-gram_brackets for PPMd. Periodic_brackets gave the worst compression times for all compression algorithms. 6-gram_chars provided the best decompression times for 7Zip, together with 2-gram_chars for BZip2, PPMd, PPMVC and 4-gram_chars for GZip and ZPAQ. Periodic_brackets provided the worst decompression times for all compression algorithms except for GZip, where prefix_chars gave the worst times.
- Overall SMILES Transform Scenarios – For all transform scenarios 7Zip provided the best compression ratios overall. PPMVC and ZPAQ gave good compression ratios. GZip gave the worst compression ratios. BZip2 and PPMd also provided

worse compression ratios. PPMd provided the best compression times. BZip2, GZip, PPMVC and ZPAQ gave good compression times. 7Zip provided the worst compression times overall. GZip provided the best decompression times overall. 7Zip and ZPAQ gave good decompression times. BZip2 and PPMd both provided worse decompression times. PPMVC provided the worst decompression times.

— Transform Comparisons – The WorstStorage SMILES transform provided the best compression ratios compared to all the other transforms tested and untransformed data when used with 7Zip. The WRT-LZ77 transform provided the worst compression ratios when used with GZip. The WRT-PAQ transform gave the best compression times when used with PPMd. The transform with the worst compression times was the WorstStorage SMILES transform when used with 7Zip. The untransformed files gave the best decompression times with GZip. The WorstStorage SMILES transform gave the worst decompression times with PPMVC.

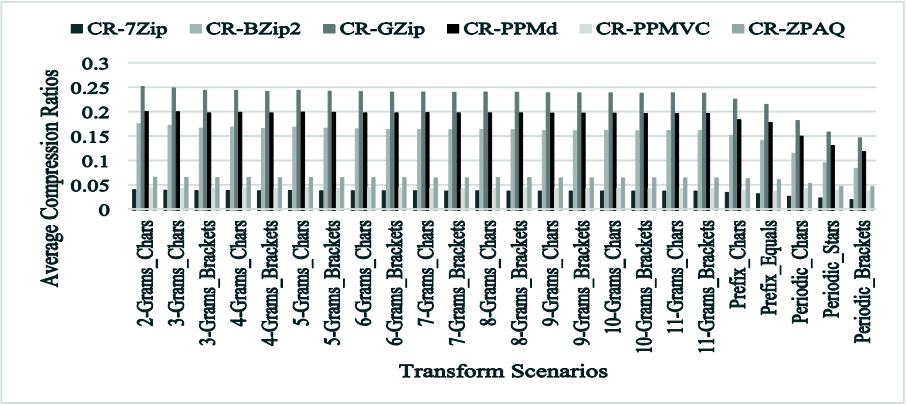


Fig. 2. Average Compression Ratios per SMILES Transform Scenarios

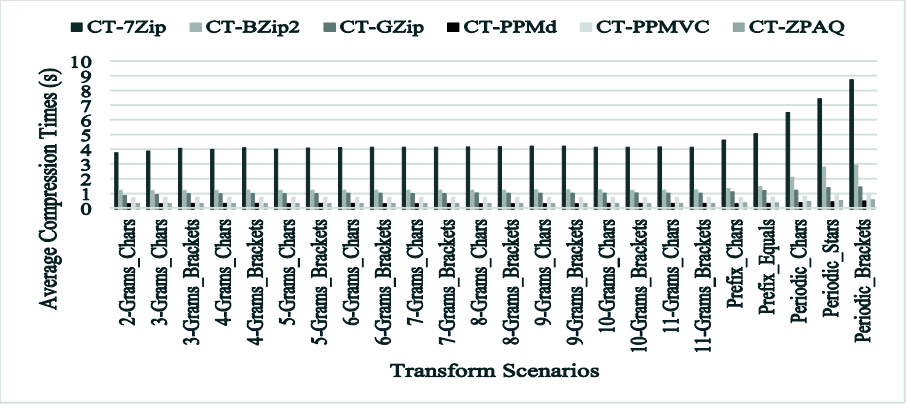


Fig. 3. Average Compression Times per SMILES Transform Scenarios

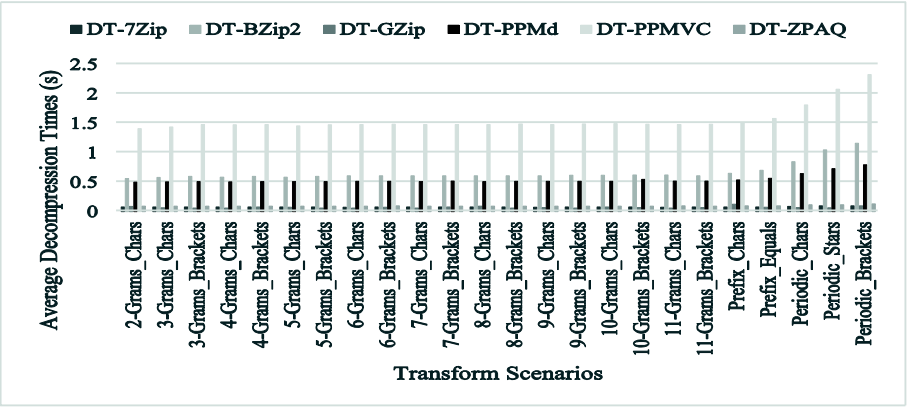


Fig. 4. Average Decompression Times per SMILES Transform Scenarios

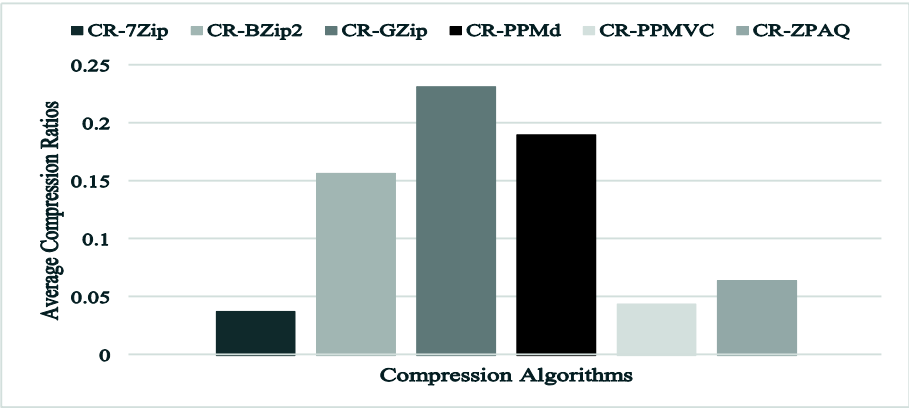


Fig. 5. Overall Average Compression Ratios for all Transform Scenarios per Algorithm

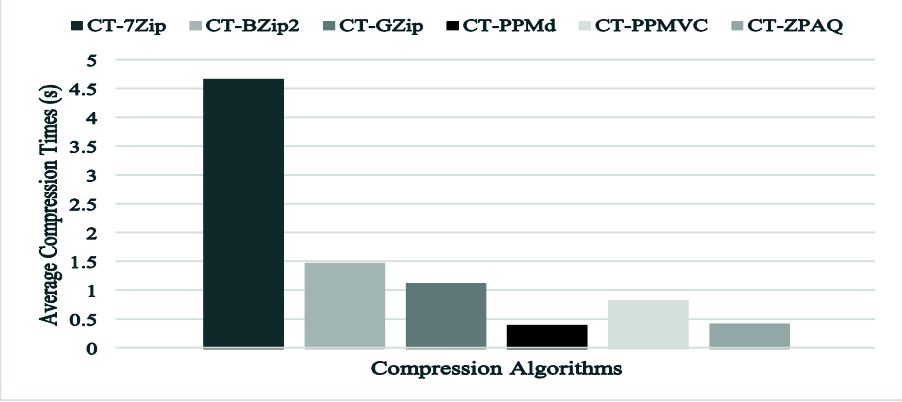


Fig. 6. Overall Average Compression Times for all Transform Scenarios per Algorithm

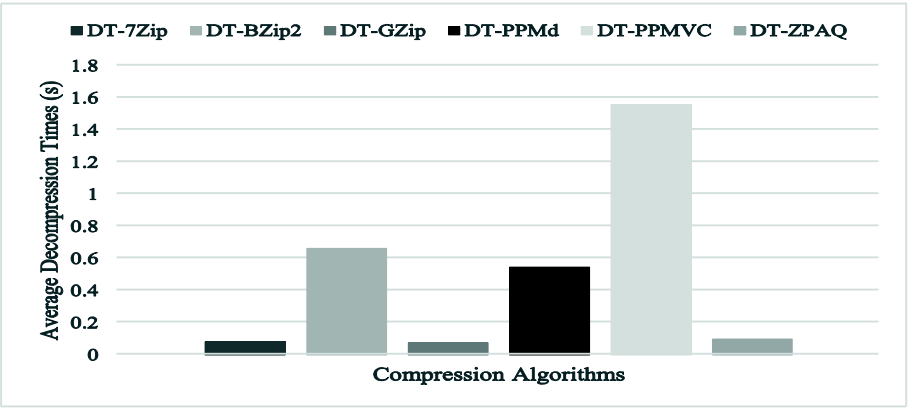


Fig. 7. Overall Average Decompression Times for all Transform Scenarios per Algorithm

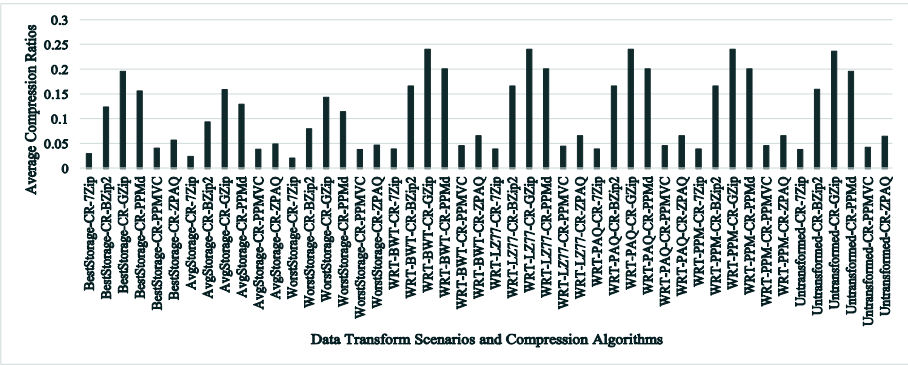


Fig. 8. Average Compression Ratios Comparing SMILES Transforms with WRT Transforms and Untransformed Data per Algorithm

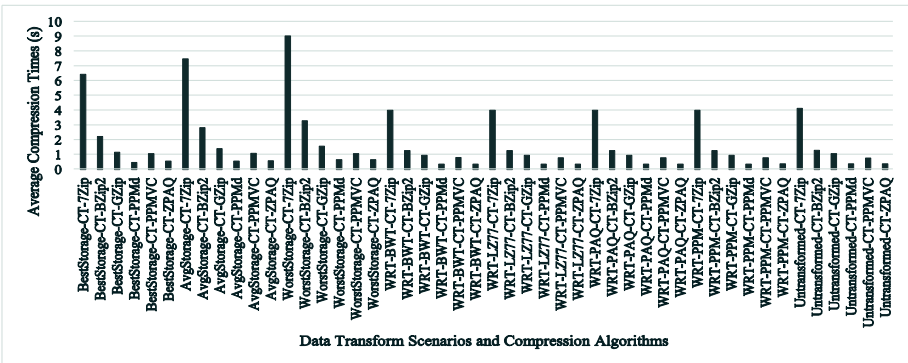


Fig. 9. Average Compression Times Comparing SMILES Transforms with WRT Transforms and Untransformed Data per Algorithm

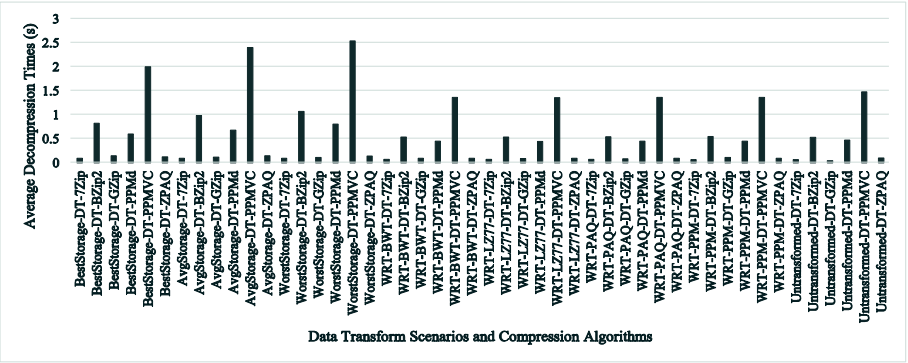


Fig. 10. Average Decompression Times Comparing SMILES Transforms with WRT Transforms and Untransformed Data per Algorithm

8 Discussions and Conclusions

In this paper we presented a SMILES-specific transform designed to enhance the compression of SMILES data when used with other general-purpose compression techniques. We tested general-purpose compression techniques on a breakdown of different SMILES transform scenarios and a combination of SMILES transforms. We compared our results with WRT transforms optimized for different compression algorithms and also on untransformed data.

The following can be concluded from the results:

- In terms of the SMILES transform scenarios tested, the results demonstrated that generally the number prefix and atomic element to atomic number transforms provided better compression ratios than n-gram substitution. However, n-gram substitution provided better compression times than number prefix and atomic element to atomic number transforms, except for PPMd. Decompression times were also generally better than the number prefix and atomic element to atomic number transforms, except for GZip and PPMd.
- It can also be noted that overall substituting data with existing characters in the data provided slightly better compression ratios than using unused characters, except for PPMVC. Whilst, many compression and decompression times for n-gram substitutions overall were better when using unused characters, the results also showed that using existing characters in some cases was better. Although, for number prefix and atomic number transforms using unused characters generally provided slightly better compression times than using existing characters. Using unused characters also provided slightly better decompression times mainly for number prefix and atomic number transforms, however, some also displayed slightly better results when using existing characters.
- The results also illustrate that generally ZPAQ provided good and balanced compression ratios, compression times and decompression times for all SMILES transform scenarios tested.

- Comparing the SMILES transforms developed with WRT optimized transforms and untransformed data, it can be concluded that all the SMILES transforms tested provided better compression ratios compared to all the WRT optimized transforms and untransformed data. Also, the WorstStorage SMILES transform provided better compression ratios than the AvgStorage and BestStorage SMILES transforms. However, all the WRT optimized transforms and untransformed data provided better compression and decompression times compared to the SMILES transforms.
- Overall, 7Zip and PPMVC provided the best compression ratios for all transforms and untransformed data. GZip gave the worst compression ratios overall. PPMd and ZPAQ provided the best compression times overall and 7Zip gave the worst compression times. 7Zip, GZip and ZPAQ provided the best decompression times, whilst PPMVC provided the worst decompression times overall.

In general, it can be concluded that transforms developed for specific data can enhance compression when used with other compressors. Transforming files prior to compression can result in larger file sizes, however, as the results have shown, compression can still be enhanced despite this. We must also not forget that transforming files incurs additional pre-processing and post-processing times prior to compression and after decompression in order to return the file back to its original state. However, in our transform it is possible to process or use the files to a certain extent without de-transforming them. This can be achieved simply because the atomic element number to atomic symbol mapping, and vice-versa, is based on the information available in the periodic table. Users can refer to the periodic table for references to the atomic number in question and process these files without de-transformation.

A final point, it is important to note when designing transform or compression techniques that providing better storage impacts compression and decompression times, as can be seen with our SMILES transforms, and vice-versa, providing better compression and decompression times can impact compression storage. It is better to try and balance storage with performance if possible.

Future work will involve further research into data transformations and compression. The transform techniques developed will be further improved, and these experiments will be extended to provide further comparisons with compression techniques and more datasets.

Acknowledgements. We would like to thank all the authors of publicly available datasets, compression and transformation tools that made this study possible. We are also very grateful to the anonymous reviewers for their valuable comments and suggestions that helped to improve this paper.

References

1. 7z Format, `http://www.7-zip.org/7z.html`
2. Benchmark Data Set for In Silico Prediction of Ames Mutagenicity, `http://doc.ml.tu-berlin.de/toxbenchmark/`
3. BZip2 for Windows, `http://gnuwin32.sourceforge.net/packages/bzip2.htm`

4. Carus, A., Mesut, A.: Fast Text Compression Using Multiple Static Dictionaries. J. Inf. Tech. 9(5), 1013–1021 (2010)
5. Cavnar, W.B., Trenkle, J.M.: N-Gram-Based Text Categorization. In: Proceedings of the Annual Symposium on Document Analysis and Information Retrieval (SDAIR 1994), Las Vegas, Nevada, USA, April 11-13, pp. 161–175 (1994)
6. Chemoinformatics.org, http://cheminformatics.org/datasets/index.shtml
7. Compression.ru Project (in Russian), http://www.compression.ru/ds/
8. Daylight Theory Manual, http://www.daylight.com/dayhtml/doc/theory/index.html
9. DSSTox, http://www.epa.gov/ncct/dsstox/
10. Engel, T.: Basic Overview of Chemoinformatics. J. Chem. Inf. Model. 46(6), 2267–2277 (2006)
11. The GZip Home Page, http://www.gzip.org/
12. Homepage of Przemysław Skibiński, http://pskibinski.pl/
13. Index of /Data/CPDB, http://www.predictive-toxicology.org/data/cpdb/
14. Karthikeyan, M., Bender, A.: Encoding and Decoding Graphical Chemical Structures as Two-Dimensional (PDF417) Barcodes. J. Chem. Inf. Model. 45(3), 572–580 (2005)
15. Kristensen, T.G., Nielsen, J., Pedersen, C.N.S.: Using Inverted Indices for Accelerating LINGO Calculations. J. Chem. Inf. Model. 51(3), 597–600 (2011)
16. Kruse, H., Mukherjee, A.: Preprocessing Text to Improve Compression Ratios. In: Proceedings of the IEEE Data Compression Conference (DCC 1998), Snowbird, Utah, USA, March 30-April 1, p. 556 (1998)
17. Mahoney, M.V.: Adaptive Weighing of Context Models for Lossless Data Compression. Technical Report CS-2005-16, Florida Institute of Technology, Melbourne, Florida, USA (2005)
18. O'Boyle, N.M.: Towards a Universal SMILES Representation – A Standard Method to Generate Canonical SMILES Based on the InChI. J. Cheminform. 4, 22 (2012)
19. Ratanaworabhan, P., Ke, J., Burtscher, M.: Fast Lossless Compression of Scientific Floating-Point Data. In: Proceedings of the IEEE Data Compression Conference (DCC 2006), Snowbird, Utah, USA, March 28-30, pp. 133–142 (2006)
20. Skibiński, P.: Reversible Data Transforms that Improve Effectiveness of Universal Lossless Data Compression. PhD Dissertation, University of Wrocław, Wrocław, Poland (2006)
21. Skibiński, P.: Two-Level Directory Based Compression. In: Proceedings of the IEEE Data Compression Conference (DCC 2005), Snowbird, Utah, USA, March 29-31, pp. 481–492 (2005)
22. Skibiński, P., Grabowski, S.: Variable-Length Contexts for PPM. In: Proceedings of the IEEE Data Compression Conference (DCC 2004), Snowbird, Utah, USA, March 23-25, pp. 409–418 (2004)
23. Weininger, D.: SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci. 28(1), 31–36 (1988)
24. ZPAQ: Open Standard Programmable Data Compression, http://mattmahoney.net/dc/zpaq.html

Personalized Web Search Using Emotional Features

Jianwei Zhang[1], Katsutoshi Minami[2], Yukiko Kawai[2], Yuhki Shiraishi[1], and Tadahiko Kumamoto[3]

[1] Tsukuba University of Technology
[2] Kyoto Sangyo University
[3] Chiba Institute of Technology
{zhangjw,yuhkis}@a.tsukuba-tech.ac.jp,
{i1258131,kawai}@cc.kyoto-su.ac.jp,
kumamoto@net.it-chiba.ac.jp

Abstract. Re-ranking and re-retrieval of search results are useful techniques for satisfying users' search intentions, since current search engines cannot always return user-desired pages at the top ranks. In this paper, we propose a system for personalized Web search considering users' emotional aspects. Given a query topic, the system presents the major emotion tendency on this topic that search results returned from search engines are reflecting. The system also enables users to specify the polarities and strengths of their emotions (e.g., happy or sad, glad or angry, peaceful or strained) on this topic and offers a re-ranking list of initial search results based on the similarity of emotions. Particularly, the system can automatically obtain Web pages with minor emotion tendency on the query topic by extracting sub-queries with opposite emotions and conducting a re-retrieval. Experimental evaluations show the re-ranking and the re-retrieval achieve encouraging search results in comparison with initial search results.

1 Introduction

Search engines such as Google and Yahoo! have become main tools to obtain information from the Web. They typically provide a list of search results for a user query, ranked by relevancy and popularity. Generally, users only look through Web pages in the top-ranked search results and may often find the information satisfying their search needs. However, current search engines sometimes fail to return user-desired pages at top ranks due to the diversity of users' search intentions. Previous works have been devoted to improving search results in two main directions: (1) re-ranking pages in the initially retrieved list and (2) suggesting query expansion or reformulation for re-retrieving new pages.

Many aspects are utilized to achieve effective re-ranking of initial search results, such as query logs [1], authorship [2], passage centrality [3], social tags [4], re-finding [5], multiple pairwise relationships between pages [6], temporal features [7], and demographical contexts (gender, age and income) [8]. Re-retrieval

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 69–83, 2013.

of new pages based on query expansion or reformulation is another effective strategy for improving retrieval accuracy, when initial search results in response to a query contain no pages relevant to users' search intentions. Query expansion or reformulation involves expanding or revising the search query to match additional or new pages by utilizing some technologies and information such as global analysis [9], pseudo-relevance feedback [10], users' personal information repository [11], term classification [12], hints obtained from external Web search engines [13], social annotation [14], wikipedia articles [15], and automatic diagnosis of term mismatch [16].

On the other hand, sentiment analysis and opinion mining [17, 18] have attracted a lot of research interests, which study emotion and its related concepts such as sentiments, opinions and attitudes. The role of emotions in information retrieval is investigated in some researches [19–21]. Specially, researches on sentiment retrieval or opinion retrieval [22–27] aim to provide a general opinion search service, similar to traditional Web search in the way that both of them find pages relevant to the query, but different from the latter in the way that sentiment retrieval need further determine whether the pages express opinions on the query topic and whether their polarities are positive or negative.

In this paper, we focus on emotional aspects of users and pages and apply such features to effectively modify search results. Not restricted to positive-negative emotions, we adopt more diverse emotions to the re-ranking and the re-retrieval of search results. Users can not only select three types of emotions ("Happy⇔Sad," "Glad⇔Angry," and "Peaceful⇔Strained."), but also set the strengths of emotions for the re-ranking. These features is also utilized to perform a re-retrieval to obtain pages with opposite emotions. Specially, we propose a system that enables the following:

- construction of emotion dictionary that represents words and their emotion values on three types of emotions.
- extraction of emotion with respect to the query topic, i.e., the major emotion tendency that search results returned from search engines are reflecting.
- interactive re-ranking of initial search results based on the user-specified emotions.
- automatic re-retrieval of new search results reflecting the opposite emotions.

Figure 1 is an example of initial search results for a query topic "Child benefit." Except for Web pages, the system also provides the emotions with respect to the query topic. Users can set their emotions and conduct a re-ranking so as to obtain a list sorted by the similarity between each initial search result's emotions and the user-specified emotions (Figure 2). Moreover, we observe that users tends to set the emotions opposite to the major emotion tendency for obtaining information in multiple perspective. Therefore, we also propose an opposite re-retrieval without user's input for automatically finding pages with minor emotion tendency (Figure 3). The query topic "Child benefit" in this example is a law introduced in Japan about distributing social security payment to the parents of children. The emotions on this topic that initial search results reflect are a little sad, a little angry and a little strained (Figure 1), because most

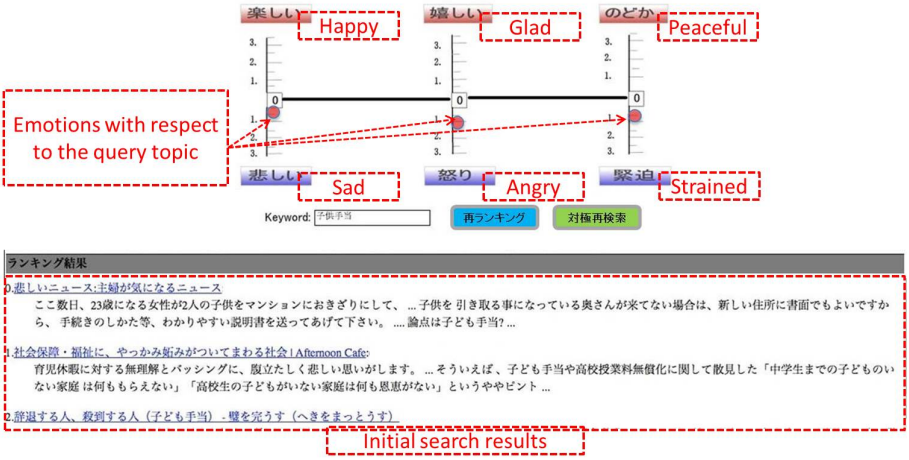


Fig. 1. Emotions with respect to the query topic

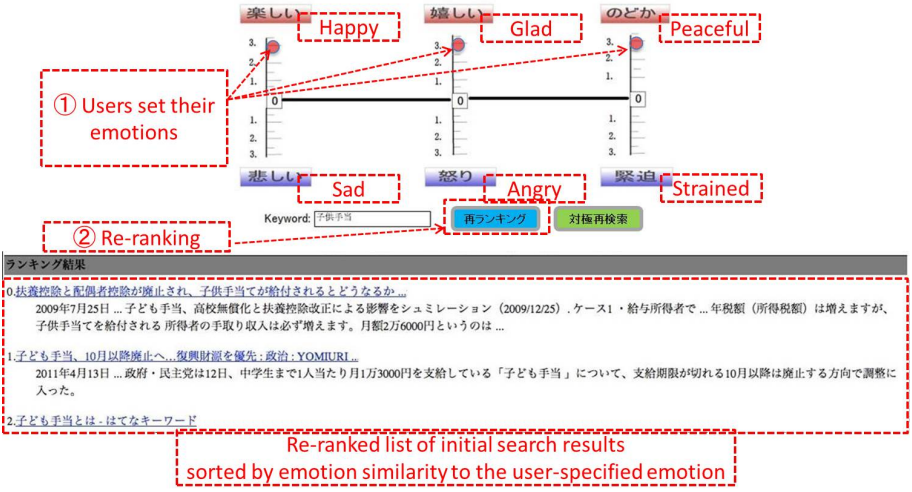


Fig. 2. Re-ranking of initial search results based on the user-specified emotion

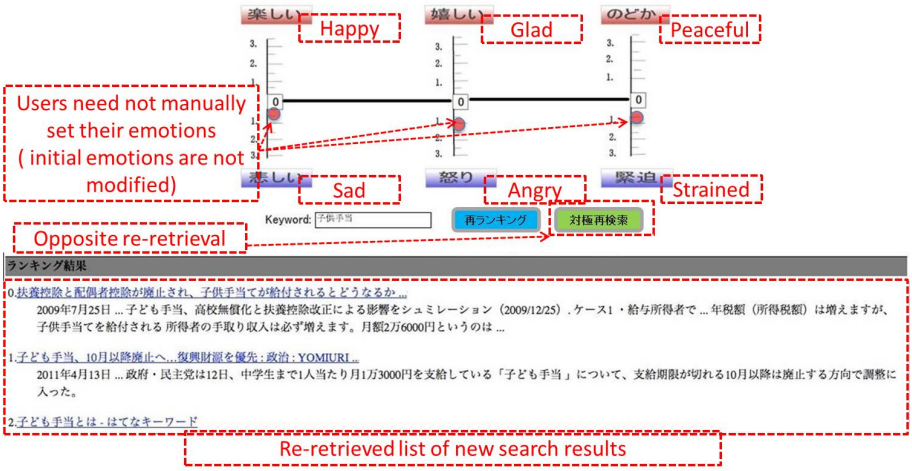


Fig. 3. Re-retrieval of new search results reflecting the opposite emotions

of pages in the initial search list introduce that many people decline this offer, or parents may abuse this grant. When a user sets the opposite emotions and conducts a re-ranking (Figure 2) or conducts an opposite re-retrieval without manually specifying any emotion (Figure 3), the pages with positive emotions can be obtained. For example, some pages at the top ranks of the re-ranking and the re-retrieval lists argue that child benefit is effective in stimulating economic growth and thus express happy, glad and peaceful emotions.

The rest of this paper is structured as follows. Section 2 provides an overview of the system. Section 3 describes the emotion calculation for search results. Section 4 describes how to modify search results using the emotional features. Section 5 evaluates the effectiveness of our system. Section 6 reviews related work. Finally we conclude the paper and discuss future work in Section 7.

2 System Overview

Figure 4 shows the overview of the proposed system. Given a query topic from a user, the system performs the following process:

1. Initial search results are returned by using Yahoo! Web Search API [28]. Specially, titles and snippets of pages are obtained for emotion analysis.
2. Emotions with respect to the query topic are calculated. The emotion values for each page ('s title and snippet) are calculated using an emotion dictionary that we have developed, and the averages of emotion values of initial search results are used as the emotions with respect to the query topic, that actually reflect the major emotion tendency of initial search results. The details of emotion calculation are described in Section 3.

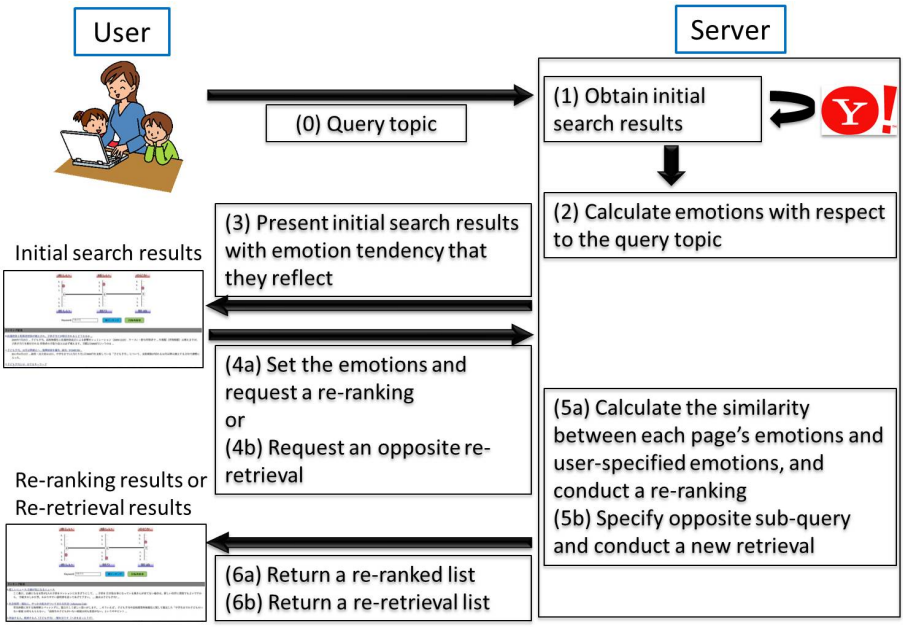


Fig. 4. Overview of the re-ranking and re-retrieval system

3. Both the initial search results and the emotion with respect to the query topic are presented to the user (Figure 1). The emotions have three dimensions and are presented as a graph.

4a. The user can set the emotions by adjusting the strength of emotion on each dimension, and request a re-ranking of initial search results based on emotional features (Figure 2). For example, the initial search results in response to the query topic "Child benefit" reflect that the major emotion tendency is "a little sad," while the user can set stronger emotion "very sad" or opposite emotion "very happy." Re-ranking is expected to sort the pages with emotions close to the user-specified one to the top ranks.

4b. In the case that the user wants to find pages with minor emotion tendency, re-ranking of the initial search results may not satisfy the user's needs if there are not the pages with opposite emotions in the initial search list. Therefore, the system enables the user to alternatively request an opposite re-retrieval (Figure 3). The re-retrieval is expected to find new pages with opposite emotions, not restricted to the initial search results. The user need not manually set the emotions for the opposite re-retrieval.

5a. For the re-ranking, the similarity between each page's emotions and the user-specified emotions is calculated. The re-ranked list of initial search results is sorted in the descending order of emotion similarity. The details of re-ranking are described in Section 4.1.

Table 1. A sample of the emotion dictionary

Word w	$s(w)$ on Happy⇔Sad	$s(w)$ on Glad⇔Angry	$s(w)$ on Peaceful⇔Strained
prize	0.862	1.000	0.808
cooking	1.000	0.653	0.881
deception	0.245	0.075	0.297
death	0.013	0.028	0.000

Table 2. Original emotion words for the three dimensions

Dimension	Original emotion words
Happy ⇔Sad	Happy, Enjoy, Enjoyment, Joy (OW_L) Sad, Grieve, Sadness, Sorrow (OW_R)
Glad ⇔Angry	Glad, Delightful, Delight (OW_L) Angry, Infuriate, Rage (OW_R)
Peaceful ⇔Strained	Peaceful, Mild, Primitive, Secure (OW_L) Tense, Eerie, Worry, Fear (OW_R)

5b. For the re-retrieval, the page with opposite emotion is first extracted from the initial search results. Then opposite sub-query extracted from this page expands the initial query to match new pages with opposite emotion. The details of re-retrieval are described in Section 4.2.

6. Finally, the re-ranking results or the re-retrieval results are presented to the user.

3 Calculation of Emotions with Respect to the Query Topic

3.1 Construction of the Emotion Dictionary

We construct an emotion dictionary, in which each entry indicates the correspondence of a word and its emotion values on three dimensions. The three-dimension emotions are "Happy⇔Sad," "Glad⇔Angry," and "Peaceful⇔Strained," that are formed based on a statistical analysis and a clustering analysis in our previous work [29]. A sample of the emotion dictionary is shown in Table 1. A emotion value $s(w)$ of a word w on each dimension is a value between 0 and 1. The values close to 1 mean the emotions of the words are close to "Happy," "Glad," or "Peaceful," while the values close to 0 mean the words' emotions are close to "Sad," "Angry," or "Strained." For example, the emotion value of the word "prize" on "Happy⇔Sad" is 0.862, which means the word "prize" conveys a "Happy" emotion. The emotion value of the word "deception" on "Glad⇔Angry" is 0.075, which means "deception" conveys an "Angry" emotion.

For each of the three dimensions, we set two opposite sets (OW_L and OW_R) of original emotion words (Table 2). The basic idea of emotion dictionary construction is that a word expressing a left emotion on a dimension often occurs with the dimension's OW_L, but rarely occurs with its OW_R. For example, the word "prize" expressing the emotion "Happy" often occurs with the words "Happy," "Enjoy," "Enjoyment," "Joy," but rarely occurs with the words "Sad," "Grieve," "Sadness," "Sorrow." We compare the co-occurrence of each target word with the two sets of original emotion words for each dimension by analyzing the news articles published by a Japanese newspaper YOMIURI ONLINE during 2002 - 2006.

First, for each dimension, we extract the set S of news articles including one or more original emotion words in OW_L or OW_R. Then, for each news article, we count the numbers of the words that are included in OW_L and in OW_R. The news articles, in which there are more words included in OW_L than in OW_R, constitute the set S_L. Inversely, the news articles, in which there are more words included in OW_R than in OW_L, constitute the set S_R. N_L and N_R represent the numbers of the news articles in S_L and S_R, respectively. For each word w occurring in the set S, we count the number of news articles including w in S_L and mark it as $N_L(w)$. Similarly, we count and mark the number of news articles including w in S_R as $N_R(w)$. The conditional probabilities are

$$P_L(w) = \frac{N_L(w)}{N_L} \qquad P_R(w) = \frac{N_R(w)}{N_R}$$

A emotion value $s(w)$ of a word w is calculated as follows:

$$s(w) = \frac{P_L(w) * weight_L}{P_L(w) * weight_L + P_R(w) * weight_R}$$

where $weight_L = log_{10}N_L$ and $weight_R = log_{10}N_R$.

3.2 Emotion Calculation for Individual Pages and Emotion Summary of Search Results

The emotion values of an individual page are calculated by looking up the emotion values of the words in the page from the emotion dictionary and averaging them[1]. In this way, a page has an emotion value ranging from 0 to 1, since the emotion values of the words in the emotion dictionary range from 0 to 1. Considering the comprehensibility and the symmetry, the emotion value (x) of a page is further converted to a value (y) ranging from -3 to 3 by the formula: $(y = 6 * x - 3)$. When x is 1, 0.5, and 0, the corresponding y becomes 3, 0, and -3. The emotion values 3, 2, 1, 0, -1, -2, -3 on a dimension, e.g., "Happy⇔Sad," correspond to "Happy," "Relatively happy," "A little happy," "Neutral," "A little sad," "Relatively sad" and "Sad" respectively.

[1] Since the title and snippet of a page summarize the content of the page and their text is shorter than full page, the system actually calculate the emotion values using the text of the title and snippet for each page so as to shorten the response time.

Emotion values on each dimension of search results in response to a query are averaged as the emotion (polarity and strength) on that dimension with respect to the query. As shown in Figure 1, Figure 2, and Figure 3, the graph has opposite emotion polarities for the three dimensions, thus represents the emotion strengths as their absolute values, although the values of negative emotions in the inner system are negative numbers.

4 Modification of Search Results Based on Emotional Features

4.1 Re-ranking of Initial Search Results Based on User-Specified Emotions

After receiving the three emotion values specified by the user, the system generates an emotion query vector $V_q = (v_{q1}, v_{q2}, v_{q3})$ using the emotion value on each dimension as its element. For each page in the initial search results list, an emotion page vector $V_p = (v_{p1}, v_{p2}, v_{p3})$ (p = 1,...,N, where N is the number of initial search results) is determined, the elements of which are the emotion values on three dimensions of each page. The similarity between V_q and V_p is calculated using the measure of cosine similarity:

$$sim(V_q, V_p) = \frac{v_{q1}v_{p1} + v_{q2}v_{p2} + v_{q3}v_{p3}}{\sqrt{{v_{q1}}^2 + {v_{q2}}^2 + {v_{q3}}^2} \times \sqrt{{v_{p1}}^2 + {v_{p2}}^2 + {v_{p3}}^2}} \quad (1)$$

The initial search results are re-ranked in a descending order of the emotion similarity and presented to the user. The pages with emotions similar to the user-specified ones tend to be ranked to the top place.

4.2 Re-retrieval of New Search Results Reflecting Opposite Emotions

The essential of the re-retrieval is to extract the sub-query to expand the initial query. For obtaining pages reflecting opposite emotions, we extract the opposite sub-query as follows:

1. The emotions on the three dimensions with respect to the initial query (emotion summary of initial search results) are represented as an emotion query vector $V_q = (v_{q1}, v_{q2}, v_{q3})$. Each page in the initial search results list is also represented as an emotion page vector $V_p = (v_{p1}, v_{p2}, v_{p3})$. The system compares the emotion polarity on each dimension between V_q and V_p, and determines the candidate opposite pages if emotion polarities on all of the three dimensions of a page are contrary to the emotion query.
2. The system then calculates the Euclidean distance $\sqrt{\sum_{i=1}^{3}(v_{qi} - v_{pi})^2}$ between V_q and V_p of each candidate opposite page. The page with the largest distance is determined as the opposite page.

Table 3. Re-ranking effect

Query topics	Emotion dimension	Emotion value of initial search results	User-specified emotion value	Emotion value after re-ranking
Child benefit	Happy⇔Sad	-0.33	0	0
	Glad⇔Angry	-1.06	3	3
	Peaceful⇔Strained	-0.77	3	3
Citizen judge	Happy⇔Sad	-0.61	3	3
	Glad⇔Angry	-0.81	3	3
	Peaceful⇔Strained	-0.63	3	3

3. Keywords are extracted from the opposite page by using Yahoo! Term Extraction API [30]. The keywords whose scores are larger than 40 are determined as the candidate sub-queries.
4. The system looks up the emotion value of each candidate sub-query from the emotion dictionary, converts the emotion values to the scale ranging from -3 to 3, and forms an emotion sub-query vector $V_s = (v_{s1}, v_{s2}, v_{s3})$. Then the system compares the Euclidean distance between V_q and each V_s. The keyword with the largest distance is determined as the opposite sub-query, and is utilized to expand the initial query for the re-retrieval.

5 Experimental Evaluation

5.1 Re-ranking Effect

We select two query topics for initial retrieval, modify the emotions of initial search results, and then conduct a re-ranking as described in Section 4.1. Table 3 shows the emotions of initial search results, the user-specified emotions, and the emotions after re-ranking. The emotion value on each dimension of initial search results is the average of emotion values of initial search results. The emotion value on each dimension after re-ranking is determined by averaging five students' evaluations on the top ten re-ranked pages. Each student read each top-ranked page and gave a level from -3 to 3 about the emotion polarity and strength that the page reflected.

From Table 3, we observe that if users set emotions different from the major emotion tendency of initial search results, the emotions after re-ranking are consistent with the user-specified emotions. This indicates that the re-ranking based on emotional features can bring the pages with emotions similar to the user-specified emotions to the top ranks.

In the experiments, we observe that users tend to set opposite emotions so as to browse different opinions. However, if there are no pages with opposite emotions in the initial search results list, it is impossible to find them by re-ranking. In that case expanding the initial query and conducting a re-retrieval are necessary for obtaining such pages. In the next section, we describe the re-retrieval effect.

Table 4. Query topics and opposite sub-queries (translated from Japanese)

Query ID	Positive Query	Opposite sub-query
(1)	Melting pot	Mixed breed
(2)	The University of Tokyo	Consumption
(3)	Internet cafe	Dispatch
(4)	April fool	Self-restraint
Query ID	Negative Query	Opposite sub-query
(5)	Press conference	Campaign
(6)	Hashimoto Toru	Schedule
(7)	Citizen judge	Plan
(8)	Tokyo Electric Power Company	Energy
Query ID	Mixed Query	Opposite sub-query
(9)	Old age	Strength
(10)	Job hunting	Media
(11)	Bubble economy	Work
(12)	Security camera	Particular

5.2 Re-retrieval Effect

We select twelve query topics for verifying the effect of the opposite re-retrieval described in Section 4.2. We compare the emotions with respect to the initial search results with the emotions with respect to the re-retrieved search results. The twelve query topics are categorized into three types based on their emotions with respect to the initial search results. If the emotions on all of the three dimensions of the initial search results are positive (Happy, Glad, and Peaceful), the query topic is categorized into Positive Query. If the emotions on all of the three dimensions of the initial search results are negative (Sad, Angry, and Strained), the query topic is categorized into Negative Query. If both positive and negative dimension exist (e.g., Happy, Glad, and Strained, etc.), the query topic is categorized into Mixed Query. Table 4 shows the query topics and their extracted opposite sub-queries.

Figure 5, Figure 6 and Figure 7 show the comparison results corresponding to each emotion dimension, respectively for positive queries, negative queries and mixed queries. The abscissa in each graph is query topics and the ordinate is emotion values. The left white bars represent the emotions with respect to the initial search results, and the right black bars represent the emotions with respect to the opposite re-retrieval search results.

From Figure 5, we can observe that for all the positive query topics, the emotion values become smaller or even change to opposite polarities. The overall average of emotion values for all the four positive query topics and all the three dimensions inclines to the negative side in a scale of 0.64. For the negative query topics in Figure 6, all except Query 7 (Citizen judge) on the dimension "Peaceful⇔Strained" obtain the emotions inclining to the positive side. The overall incline scale from the negative side to the positive side is 0.54. Opposite re-retrieval is also effective for the mixed query topics in Figure 7. Except Query

11 (Bubble economy) on the dimensions "Happy⇔Sad" and "Glad⇔Angry," opposite re-retrieval succeeds to obtain search results with emotions inclining to the reverse side for all other mixed query topics. The overall incline scale (from negative to positive, or from positive to negative) for the mixed query topics is 0.14.

There are some comprehensible pages for comparing the contents of the initial and re-retrieved search results. For example, for Query 4 (April fool), the initial search results contain the pages that introduce tricks for April fool and express happy emotions, while the pages in the top re-retrieved search results are the sad pages about refraining from tricks. Another comprehensible example is Query 8 (Tokyo Electric Power Company). The pages in the initial search list in response to this query are the negative pages expressing considerable criticism to the way that TEPCO handled the crisis of nuclear accidents, while opposite re-retrieval brings the positive pages that introduce TEPCO's efforts to develop future energy.

6 Related Work

There have been a number of studies on re-ranking search results considering various aspects. Zhuang et al. [1] proposed a Q-Rank method to refine the ranking of search results by constructing the query context from query logs. Bogers et al. [2] utilized the authorship information to extract expert rankings. Bendersky et al. [3] presented a passage-based approach to leverage information about the centrality of the document passages with respect to the initial search results list. Yan et al. [4] proposed a Query-Tag-Gap algorithm to re-rank search results based on the gap between search queries and social tags. Tyler et al. [5] utilized the prediction of re-finding (finding the pages that users have previously visited) to re-rank pages. Kang et al. [6] proposed to combine multiple pairwise relationships between documents to re-rank search results. Chang et al. [7] exploited temporal features for re-ranking time-sensitive search results. Kharitonov et al. [8] improved search results by using demographical contexts such as gender, age, and income.

Another research direction for improving retrieval accuracy is to re-retrieve new search results based on query expansion or reformulation. Xu et al. [9] showed that using global analysis such as word context and phrase structure on local documents (the documents retrieved by the initial query) produced more effective search results. Tao et al. [10] proposed to integrate the initial query with pseudo-relevance feedback documents in a probabilistic mixture model without parameter tuning. Chirita et al. [11] expanded short queries by analyzing user data at different levels ranging from term and compound level analysis up to global co-occurrence statistics. Cao et al. [12] argued against the assumption of pseudo-relevance feedback that the most frequent terms in the pseudo-feedback documents are useful for the retrieval, and proposed to integrate a term classification process to predict the usefulness of expansion terms. Yin et al. [13] utilized the hints such as query logs, snippets and search result documents from

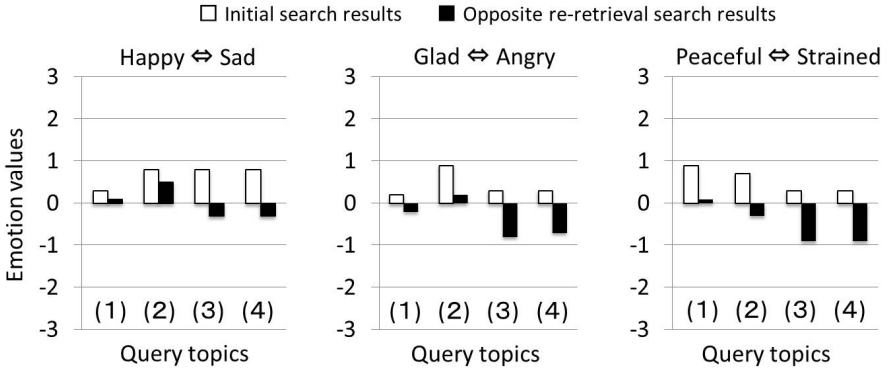


Fig. 5. Positive query

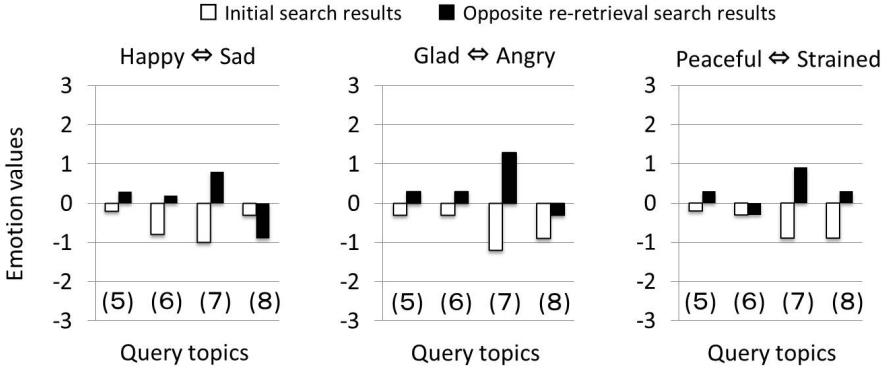


Fig. 6. Negative query

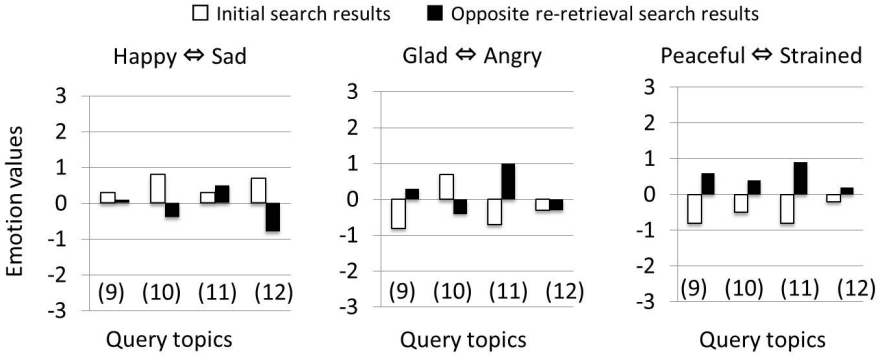


Fig. 7. Mixed query

external search engines to expand the initial query. Lin et al. [14] extracted candidate expansion terms by a term-dependency method and ranked them based on social annotation resource. Oliveira et al. [15] proposed to expand entity-related queries using wikipedia articles and tag recommendation methods. Zhao et al. [16] proposed an automatic diagnosis of term mismatch to guide interactive query expansion or create conjunctive queries.

Similar to these researches we also aim to obtain a user-desired re-ranking and re-retrieval search results. Different from them we improve a personalized Web search using the emotional features. The pages with emotions similar to users' can be re-ranked to the top rank and the pages with opposite emotions can be re-retrieved based on the extraction of opposite sub-queries.

On the other hand, sentiment analysis and opinion mining [17, 18] are one of the hottest research areas that extract sentiments (or emotion, opinions, attitudes) from text such as movie reviews, book reviews, and product evaluations. Some researches have applied emotion knowledge to information retrieval and its relevant research areas. Arapakis et al. [19] found that emotions in the information seeking process interwove with physiological, psychological and cognitive processes and formed characterized patterns according to specific task and specific user. In [20], Arapakis et al. personalized affective models using facial expression data and compared their performance to a general model for predicting topical relevance. Moshfeghi et al. [21] investigated the role of emotional features in collaborative recommendation and their experimental results showed the emotional features extracted from movie reviews were capable of enhancing recommendation effectiveness.

Specially, sentiment retrieval or opinion retrieval is a newly developed research subject, which requires documents to be retrieved and ranked according the opinions about a query topic. Eguchi et al. [22] proposed several sentiment retrieval models based on probabilistic language models, assuming that users both input query topics and specify sentiment polarity. Similar methods proposed in [23] and [24] unified topic relevance and opinion relevance respectively based on a quadratic combination and a linear combination. A different method proposed by Li et al. [25] combined topic-sentiment word pairs in a bipartite graph to effectively rank the documents. Opinion retrieval from UGC (User Generated Content) such as blogs [26] and Twitter [27] also yields comparable retrieval performance.

Different the above researches mainly focusing on review documents and positive-negative emotions, we consider any Web pages and more diverse emotions. As we showed in the experiments, there are Web pages that express both positive emotion and negative emotion in different dimensions.

7 Conclusions and Future Work

In this paper, we studied the problem of personalized Web search considering emotional aspects: re-ranking initial search results based on user-specified emotions, and re-retrieving new search results with opposite emotions based on

the extraction of opposite sub-query. Given a query topic, the system shows the major emotion tendency that search results reflect. After users specify their emotions, the search results similar to users' emotions can be re-ranked to the top place. Even without users' interaction, the system also automatically extracts opposite sub-query so as to re-retrieve the pages with minor emotion tendency with respect to the initial query. We have showed that our system is effective in re-ranking the initial search list and re-retrieving opposite pages.

In the future, we plan to repeat the re-retrieval process and verify whether Web pages with stronger opposite emotions can be extracted. In that case, the problems about the reversion of emotion polarity (from positive to negative, then back to positive) and the topic drift must be taken into consideration.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number 24800007, 23500140, 24500134, 24780248.

References

1. Zhuang, Z., Cucerzan, S.: Re-ranking search results using query logs. In: CIKM 2006, pp. 860–861 (2006)
2. Bogers, T., van den Bosch, A.: Authoritative re-ranking of search results. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 519–522. Springer, Heidelberg (2006)
3. Bendersky, M., Kurland, O.: Re-ranking search results using document-passage graphs. In: SIGIR 2008, pp. 853–854 (2008)
4. Yan, J., Liu, N., Chang, E.Q., Ji, L., Chen, Z.: Search result re-ranking based on gap between search queries and social tags. In: WWW 2009, pp. 1197–1198 (2009)
5. Tyler, S.K., Wang, J., Zhang, Y.: Utilizing re-finding for personalized information retrieval. In: CIKM 2010, pp. 1469–1472 (2010)
6. Kang, C., Wang, X., Chen, J., Liao, C., Chang, Y., Tseng, B.L., Zheng, Z.: Learning to re-rank Web search results with multiple pairwise features. In: WSDM 2011, pp. 735–744 (2011)
7. Chang, P., Huang, Y., Yang, C., Lin, S., Cheng, P.: Learning-based time-sensitive re-ranking for Web search. In: SIGIR 2012, pp. 1101–1102 (2012)
8. Kharitonov, E., Serdyukov, P.: Demographic context in Web search re-ranking. In: CIKM 2012, pp. 2555–2558 (2012)
9. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: SIGIR 1996, pp. 4–11 (1996)
10. Tao, T., Zhai, C.: Regularized estimation of mixture models for robust pseudo-relevance feedback. In: SIGIR 2006, pp. 162–169 (2006)
11. Chirita, P., Firan, C.S., Nejdl, W.: Personalized query expansion for the Web. In: SIGIR 2007, pp. 7–14 (2007)
12. Cao, G., Nie, J., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: SIGIR 2008, pp. 243–250 (2008)
13. Yin, Z., Shokouhi, M., Craswell, N.: Query expansion using external evidence. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 362–374. Springer, Heidelberg (2009)
14. Lin, Y., Lin, H., Jin, S., Ye, Z.: Social annotation in query expansion: a machine learning approach. In: SIGIR 2011, pp. 405–414 (2011)

15. Oliveira, V., Gomes, G., Belem, F., Brandao, W.C., Almeida, J.M., Ziviani, N., Goncalves, M.A.: Automatic query expansion based on tag recommendation. In: CIKM 2012, pp. 1985–1989 (2012)
16. Zhao, L., Callan, J.: Automatic term mismatch diagnosis for selective query expansion. In: SIGIR 2012, pp. 515–524 (2012)
17. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2007)
18. liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)
19. Arapakis, I., Jose, J.M., Gray, P.D.: Affective feedback: an investigation into the role of emotions in the information seeking process. In: SIGIR 2008, pp. 395–402 (2008)
20. Arapakis, I., Athanasakos, K., Jose, J.M.: A comparison of general vs personalised affective models for the prediction of topical relevance. In: SIGIR 2010, pp. 371–378 (2010)
21. Moshfeghi, Y., Jose, J.M.: Role of emotional features in collaborative recommendation. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 738–742. Springer, Heidelberg (2011)
22. Eguchi, K., Lavrenko, V.: Sentiment retrieval using generative models. In: EMNLP 2006, pp. 345–354 (2006)
23. Zhang, M., Ye, X.: A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In: SIGIR 2008, pp. 411–418 (2008)
24. Huang, X., Croft, W.B.: A unified relevance model for opinion retrieval. In: CIKM 2009, pp. 947–956 (2009)
25. Li, B., Zhou, L., Feng, S., Wong, K.: A unified graph model for sentence-Based opinion retrieval. In: ACL 2010, pp. 1367–1375 (2010)
26. Zhang, W., Yu, C.T., Meng, W.: Opinion retrieval from blogs. In: CIKM 2007, pp. 831–840 (2007)
27. Luo, Z., Osborne, M., Wang, T.: Opinion retrieval in Twitter. In: ICWSM (2012)
28. Yahoo! Web Search API, `http://developer.yahoo.co.jp/webapi/search/websearch/v2/websearch.html`
29. Zhang, J., Kawai, Y., Kumamoto, T., Nakajima, S., Shiraishi, Y.: Diverse sentiment comparison of news websites over time. In: Jezic, G., Kusek, M., Nguyen, N.-T., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2012. LNCS, vol. 7327, pp. 434–443. Springer, Heidelberg (2012)
30. Yahoo! Term Extraction, `http://developer.yahoo.com/search/content/V1/termExtraction.html`

MetaExtractor: A System for Metadata Extraction from Structured Data Sources

Alexandra Pomares-Quimbaya, Miguel Eduardo Torres-Moreno, and Fabián Roldán

Pontificia Universidad Javeriana, Bogotá, Colombia

Abstract. The extraction of metadata used during the planning phase in mediation systems assumes the existence of a metadata repository that in most cases must be created with high human involvement. This dependency rises complexity of maintenance of the system and therefore the reliability of the metadata itself. This article presents MetaExtractor, a system which extracts structure, quality, capability and content metadata of structured data sources available on a mediation system. MetaExtractor is designed as a Multi-Agent System(MAS) where each agent specializes in the extraction of a particular type of metadata. The MAS cooperation capability allows the creation and maintenance of the metadata repository. MetaExtractor is useful to reduce the number of data sources selected during query planning in large scale mediation systems due to its ability to prioritize data sources that better contribute to answer a query. The work reported in this paper presents the general architecture of MetaExtractor and emphasizes on the extraction logic of content metadata and the strategy used to prioritize data sources accordingly to a given query.

Keywords: Large Scale Data Mediation, Metadata Extraction, Source Selection, Multi-Agent System, Mediation Systems.

1 Introduction

The processing of queries in mediation systems involves processes of logic and physical planning taking into account the characteristics of the data sources available in the system. This information is typically stored in metadata repositories that are accessed whenever a query is evaluated. Unfortunately, mediation systems assume the existence of previously constructed metadata repositories [1] or focuses only on the structure of data sources, and basic statistics such as the number of records, the number of null data, etc.[2] Although advanced mediation strategies contemplate the existence of more comprehensive metadata such as the number of objects contained in a source [3], the way in which this metadata is obtained has not been well defined. Parallel to mediation systems of structured sources, several proposals have been made to summarize the contents of available sources in the Web: using trees [4] to describe the dependency relationships between sources [5]. These alternatives are useful for increasing the

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 84–99, 2013.

richness of metadata. However, their generation requires high human interaction or includes only general aspects that do not allow a real differentiation among sources, especially when the sources are dependent and data fragmentation is not disjoint. Faced with these limitations of mediation systems this article proposes MetaExtractor a system designed for the automatic extraction of metadata from structured distributed data sources that are registered in a mediation system. MetaExtractor's goal is to increase the quality of metadata repositories through methods that do not require intensive human involvement. MetaExtractor design is based on multi-agent systems that uses data sources as resources to meet its goal of extracting a particular type of metadata. MetaExtractor considers the extraction of structure, capability, quality and content metadata. The structure of this article presents in Section 2 the analysis of related works on the area of mediation systems that allow to put into context the contribution of MetaExtractor. Subsequently, Section 3 presents an overview of MetaExtractor that is detailed in Section 4. Section 5 reports evaluation results on priorization of data sources using the extracted metadata. Finally, Section 6 presents the conclusions and future work.

2 Related Work

Different approaches to support data queries in the context of heterogeneous and distributed sources have been studied for over 15 years. Most of them are based on mediation systems [6] that aims to provide an integrated view of heterogeneous and distributed data sources respecting their own autonomy [7]. The architecture of mediation systems [7,8,9,10] consists of four levels of abstraction: i) the data sources level, ii) the adaptation (wrappers) level, iii) the mediation level and iv) the application level. Query processing in this type of systems is coordinated at the mediation level and takes place in two stages: planning and execution. In the planning stage, sources that can fully or partially answer the query are selected and then the query is rewritten into subqueries, using the external model of each source. To select appropriate sources, the mediator uses a metadata catalogue that includes the external model and other information sources, whose level of detail and accuracy varies according to the type of mediator. In the execution stage, the mediator sends the subqueries to each source and is responsible for coordinating the execution of the queries, and receiving and processing responses in order to integrate them.

One factor that affects the accuracy and efficiency of a mediation system is its planning strategy, which is directly related to the quantity and quality of available metadata. In the first generation of mediation systems, such as: Information Manifold [11], TSIMMIS [9], DISCO [8] and Minicon [12], the planning process uses as metadata information the processing capabilities of sources (e.g. the number of required conditions, or the attributes that could be defined in the predicate). Another group of strategies such as: iDrips and Streamer [1], navigation paths [3] perform the planning process by means of metadata that describes the content of the sources. Other proposals such as QPIAD [13] and

the quality-oriented presented in [14] use detailed statistics on the quality of the data contained in the sources to reduce the number of plans. Meanwhile, proposals for large-scale mediation: PIER [15], PinS [16], Piazza [17], PeerDB [18], and SomeWhere [19] use location of metadata, content summary and description of reputation sources, among others.

The metadata used in these systems are an essential input for the performance of the mediation process. However, obtaining such metadata from structured data sources is limited to the manual (human) description of the structure of the sources and some quality statistics obtained during the operation of the system. Even if the owners or managers of each source of information can define some metadata manually, to ensure maintainability, efficiency, consistency and cost of the mediation system, it is necessary to generate strategies for automatically obtain metadata from the mediation repositories.

The existing methods for generating metadata can be classified into three groups: i) extraction, ii) collection and iii) hybrid generation [20]. In the first case the aim is to obtain the metadata based on the execution of techniques for analyzing, mining, evaluating, etc. In the case of metadata collection the goal is to gather pre-built metadata. Finally, hybrid generation obtains existing metadata, and from its analysis generates new metadata [21]. The aim of the MetaExtractor project is focused on metadata extraction from structured data sources that may contain narrative text in some of their attributes.

Literature reviews show that most research efforts suggests strategies to extract metadata from unstructured information available on the web, the vast majority of available methods use processing algorithms that analyze natural language texts in order to generate structured descriptions [22]. These algorithms are part of strategies for information retrieval that assess text in a semantic and syntactic way to get a list of descriptors for the content of the sources [23]. Although there are methods of metadata generation from structured sources its extraction and / or collection is focused on basic attributes such as size or scheme that cannot act as a true differentiator of data sources.

This article presents MetaExtractor a system that modifies and extends these existing methods and techniques to enhance the extraction of metadata from structured sources. MetaExtractor obtains metadata using different techniques according to the type of metadata that must be extracted. In addition, given a query, MetaExtractor prioritizes data sources that better contribute to answer it. The proposed MetaExtractor architecture includes the use of data mining techniques to identify the type of content from a source. The intention is to explore a sample of each data source to establish whether there is a tendency to have some type of instances of a business object (e.g. client, patient).

3 MetaExtractor Overview

As Section 2 illustrates, current mediation system strategies for source selection assume the existence of the required metadata for query planning, extracts only basic metadata, or require an important manual effort to feed metadata repositories. In order to fulfill this gap this paper proposes MetaExtractor a system

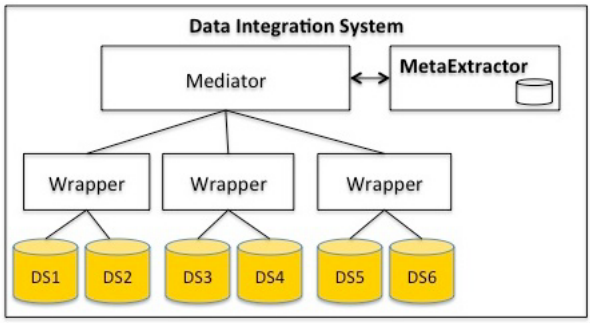


Fig. 1. MetaExtractor in a Mediation System

that allows to extract structure, quality, capability and content metadata from structured data sources. MetaExtractor architecture is based on the interaction of software agents in charge of extracting different types of metadata from distributed and heterogeneous structured data sources. The main idea behind the agent architecture is for MetaExtractor to have agents located in the same location as the data source and watching over the sources' logs to identify any important or relevant changes on the structure or data contained in each source. This section presents an overview of MetaExtractor functionality and architecture. The following section emphasizes on key aspects of the agents in charge of metadata extraction.

3.1 MetaExtractor Preliminaries

The goal of MetaExtractor is to provide the metadata required for query planning in a mediation system. Figure 1 illustrates the relationship between MetaExtractor and a mediation system, the goal is to act as the knowledge base during the planning phase of a query.

MetaExtractor extracts and stores metadata based on the assumption that data sources contain pieces of information of the relevant objects in a specific domain (e.g. patient, client, campaign) from now on called VDO (see Definition 1). The VDO illustrated in Figure 2 represents a *Patient* composed by four concepts (Person, Medical Act, Medical History, Affiliation). This VDO pertains to the medical domain, which is going to be used from now on to describe MetaExtractor's functionality.

Definition 1 ***Virtual Data Object (VDO)****. A VDO is a logical set of related concepts relevant to a group of users. Figure 2 illustrates a VDO joining four concepts. Each concept has data type properties whose values are data literals. Concepts are related through object type properties whose values are instances. VDOs are virtual because their instances are partitioned and distributed on several data sources.*

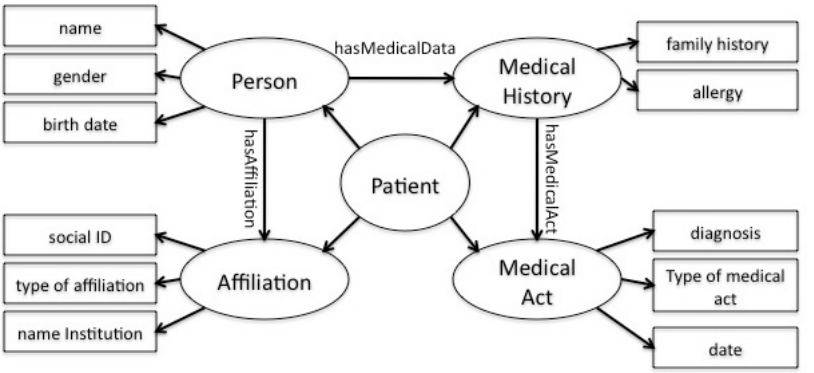


Fig. 2. VDO Patient

MetaExtractor considers queries with selection conditions on the properties of the concepts composing the VDO. For example, the query hereafter uses the *VDO Patient* and selects the id of patients with a "Buphtalmos" diagnosis.

Query 1 $Q(VDO_{Patient}, properties(id), conditions(diagnosis = Buphtalmos))$

In the following sections we will work with queries which include several conditions, such as follows:

$Q(VDO_{name}, properties(id), conditions(condition_1, condition_2, ..., condition_n))$

When a mediator receives a query it can issue a request to MetaExtractor asking about the facts related to the VDO involved in the query; these facts may be given directly to the mediator or they can be used by MetaExtractor to prioritize the data sources that better contribute to answer the query.

3.2 MetaExtractor Processes

Figure 3 illustrates the agents and processes involved in the execution of MetaExtractor. The GUI process, the Server process and the Source process include a set of sub-processes (agents), which corresponds to activities of an agent with a defined behaviour. The GUI agents represent the interactions of the users, the Server agents manage persistence of metadata and users, and include the logic to prioritize data sources given a particular query. To accomplish their tasks they use the metadata extracted by the Source agents that are in charge of the actual extraction of metadata and are the ones watching over the source itself.

The metadata extraction starts with a communicative act between the Metadata Manager Agent, who proactively and periodically activates the metadata extraction. The activation event triggers the activation of the Source Manager Agent and this, in turn, sends an event to the related Data Source Agents. Finally, each Data Source Agent communicates with the Extraction Coordination

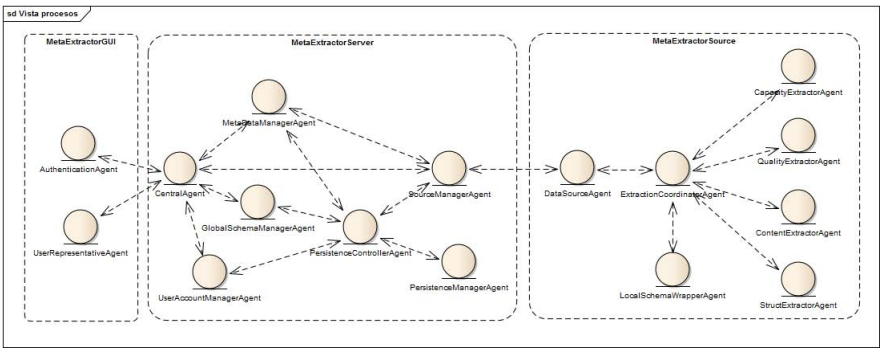


Fig. 3. MetaExtractor Agents Interaction

Agent that is in charge of coordinating the extraction of each type of metadata. Currently, MetaExtractor provides four agents in charge of extracting diferent metadata types: structure, quality, content and capability metadata. The following subsection describes each one of metadata types and Section 4 the internal logic of each metadata extractor agent. It is important to notice that according to the requirements of the mediation system, different kinds of metadata extractor agents can be added later on to the system.

Extraction Functions. Once a data source is registered MetaExtractor is able to extract and/or obtain from them the following metadata (by means of the Source Agent):

1. Structure Metadata: This type of metadata describes the relationship between the objects of a data source with the domain VDOs. For instance, a data source in the medical domain may have a table called *Attention* that contains all the attentions of a patient in a medical institution, this table may be the equivalent of *Medical Act* in the description of the VDO patient. In this case the metadata structure contains the equivalent relationship between the table *Attention* and the concept *Medical Act*. Structure metadata includes the name of the tables, attributes and relationships, and their relationship with the domain VDOs by using the operators Equivalence and Inclusion.
2. Quality Metadata: MetaExtractor provides information about the quality state of a data source. This type of metadata is very important to distinguish which data source is more promising to answer a query when two or more data sources have similar structural and content characteristics. The quality state is described using the following attributes:
   - Null percentage: It serves to identify the level of deterioration of an attribute according to the frequency of nulls related to the size of a sample. An attribute with a value 65% in Null Percentage can mean that the quality is too low.

  - Null density: It calculates the density of null values in a table, its goal is to differentiate tables that have few attributes with high frequency of null values and tables that have a larger number of attributes with low frequency of null values.
  - Maximum null frequency: Identify the attribute with the highest frequency of nulls.
3. Content Metadata Extraction: Most of the queries over mediation systems involve a condition over a set of attributes of the VDO; for instance in the medical domain, a query over the patient VDO may have a condition over the diagnosis (e.g. diagnosis = cancer). These conditions include a restriction of value over an attribute. The content metadata allows to identify the data sources that better could answer a query according to the role it can play as a contributor of instances of a VDO in the domain where the restriction is true. Roles reflect the ability of data sources to resolve query predicates. Section 4 describes the roles and the strategy used by MetaExtractor to describe the content of a data source.
4. Capability Metadata Extraction: This type of metadata is extracted to identify data sources that have restrictions over the type of queries that they can answer. This metadata is specially important in data sources that can only be accessed through a set of services or an interface that limits the attributes that may be acquired. It is also relevant when data sources demand a specific number of query conditions to answer a query. In the medical domain for instance, a data source may restrict the access of the patient's name or may demand the inclusion of at least one condition in the query.

MetaExtractor provides three functions for querying the metadata required for planning purposes. These functions are:

- Query data source metadata: Allows to query the values of metadata of a specific data source.
- Prioritize data sources: According to the properties and conditions related to a VDO query it delivers the list of prioritized data sources to answer it. In order to prioritize MetaExtractor uses the logic proposed in [24]. This strategy is based on combinatorial optimization techniques.
- Metadata subscription: Allows to obtain changes on the metadata of a data source or a set of data sources automatically.

4 Extractor Agents

This section presents the logic within content, quality and capability metadata extractor agents.

4.1 Content Metadata Extractor Agent

The goal of this agent is to identify the role that a data source may play during the execution of a query taking into account its content. This task is not

straightforward in structured data sources, because they may contain different records, each one of them containing different content. As a consequence, the *Content Metadata Extractor Agent* must identify the better way to describe a data source given the content of its records. Lets look an example. Consider three hospital databases DS1, DS2 and DS3. According to the kind of hospital these databases may be specialized on a specific type of patients. For example DS1 is specialized on information about patients with cancer. DS2 is specialized on pediatrics and, as a consequence, it contains only child patients. And DS3 contains both, patients with cancer, child patients, as well as other type of patients because is the database of a general hospital. In this domain the agent must describe the role according to the attribute "diagnosis" and "age" differentiating the specialization of DS1 (i.e. diagnosis = Cancer) and DS2 (age < 18), and indicating that even if DS3 is not a specialist it may contain records with this content.

Roles reflect the ability of sources to solve conditions. Given the analysis of the roles played by a database in a mediation system, we propose the following roles: *specialist* and *container*. The definition of each source role is described in Definition 2 and 3. In these definitions all the instances of *VDOj* stored in data source *DSi* are noted $ext(DSi, VDOj)$ [1]. U designates all the data sources participating in the system. All the instances of *VDOj* available in the system are denoted $ext(U, VDOj)$. The subset of $ext(DSk, VDOj)$ corresponding to the instance that verifies a condition p is denoted $ext(DSk, VDOj)^p$ and $card()$ is the cardinality function.

Definition 2 ***Specialist Role.*** *A data source DSi plays a specialist role w.r.t. a condition p in a query on VDOj iff most instances of VDOj stored in DSi match p.* $IsSpecialist(DSi, VDOj, p) \Longrightarrow card(ext(DSi, VDOj)^p) \geq card(ext(DSi, VDOj)^{\neg p})$

Definition 3 ***Container Role.*** *A data source DSi plays a container role w.r.t. a condition p in a query on VDOj iff DSi contains at least one instance of VDOj that matches p.* $IsContainer(DSi, VDOj, p) \Longrightarrow ext(U, VDOj)^p \cap ext(DSi, VDOj)^p \neq \varnothing$

The process of content metadata extraction is divided into two phases, the Domain Training phase and the Content Description. The following subsections describe both of them.

Domain Training. The content of a data source must be described in terms of its domain. For instance, a data source containing electronic health records (EHR) must be described in terms of diagnosis, treatments, medicines, or in terms of the properties of patients like age and gender. Typically, a data source of EHR contains attributes that store all this information; however, some of them, and frequently the most important, are narrative text attributes that contain these terms hidden in lots of text.

The goal of this phase is to record and train the system to recognize what kind of VDOs are contained in a data source. In order to do that MetaExtractor

[1] This extension contains the VDO identifiers.

must be trained to identify when the content of a data source is related to a VDO with a specific value in one or more properties. For instance, when a data source contains patients with a specific diagnosis. This identification is straightforward when data sources contain mostly well coded attributes; however, as mentioned, in some domains (e.g. medical) this information may be within texts.

This phase works as follows, for each one of the VDO properties that are usually used as query conditions (e.g. diagnosis) and their possible values according to well known thesaurus or terminology dictionaries (e.g. SNOMED CT [32]), MetaExtractor searches on the internet (using an API created for this purpose) in order to identify what are the common words used when talking about the condition with a specific term (e.g. diagnosis and diabetes). The search involves the property and its possible terms (e.g. disease, diabetes). According to the domain of the mediation system, the properties, as well as their possible terms, must be parametrized in the system. At the end of this training phase each one of the terms related to selected properties has associated a dictionary of related words used in conjunction with the term. The outputs are called training dictionaries, and are used later to identify the terms that better describe a data source.

Content Description. This phase identifies the role played by a data source within the system. The algorithm proposed to identify the role is illustrated in Algorithm 1. The principle of this algorithm is to use narrative text attributes and a subset of structured attributes from the data source to identify its main content. For example, the narrative text attributes of an EHR that contain the evolution and the treatment of the patient are very important to describe the type of patients the data source has. Additionally, some of the structured attributes (i.e. attribute that contains a limited number of possible values) may provide hints on the content of the data source.

In case the size of a data source exceeds a threshold, the first part of the algorithm reduces its size applying a stratified sample technique [I]. This technique uses a structured attribute to identify the strata.

Then, for each one of the records in the data source sample [II], and using text mining techniques, it annotates the narrative texts taking into account the training dictionaries created in the *Domain Training* phase [III]. As an example we can obtain for a record the annotation (diagnosis = heart failure, sign = pulmonary edema).

Next, analyzing the structured attributes it identifies its value and adds it to the annotations [IV]; After this step the annotation may include (diagnosis = heart failure, sign = pulmonary edema, gender = F).

Once MetaExtractor analyzes all the records it applies a neighborhood based clustering algorithm [25] to identify if there is a specific group of similar records [V] in the data source. The annotations that better describe each cluster are used to specify the role of the data source. In case a cluster contains more than 50% of the records of the data source, the data source will be stored as a specialist with respect to the property with the value that better describes the cluster (e.g. diagnosis:heart failure) [VI]; otherwise, it will be stored as a container.

Algorithm 1. Role Identification

```
Input:  DS //DataSource
        Cat(cat1{v1,...,vn},...,catm{v1,...,vn})
        sizeThreshold
Output: QRoles(role(catj),...,role(catk))
Begin
    QRoles={}, annotations={}, clusters={}
[I] If (size(DS) > sizeThreshold)
       s = getSample(DS)
     Else
        s = DS
[II] ForAll (reg in s)
       ForAll (att in reg)
          If (narrativeAtt(att))
[III]             annotations[reg] = annotations[reg] + annotate(att)
          Else
[IV]              annotations[reg] = annotations[reg] + idValue(att)
[V] clusters = Cluster(annotations)
    ForAll (c in clusters)
[VI]       If (size(c) > size(s)/2)
            QRoles = QRoles(specialist(c.cat))
       Else
            QRoles = QRoles(container(c.cat))
   Return(QRoles)
End
```

4.2 Quality Metadata Extractor Agent

This agent evaluates the quality state of a data source based on the analysis of nullity of its attributes. To obtain this state MetaExtractor evaluates the frequency of null values in each one of the attributes and the general density of nullity of the data source. The density is very important because some attributes may have high frequency of nullity not because of the quality, but because of the domain. For instance, an attribute middle name may have a lot of nulls because is an optional value in the domain, its null frequency does not reveal a quality problem, but a business rule. In order to assure the performance of MetaExtractor when data sources contain high number of records, it provides the possibility of obtaining quality metadata over a sample from the original data source. This sample follows a systematic approach using as sample size a parameter that can be configured in the system.

4.3 Capability Metadata Extractor Agent

Capability metadata extractor goal is to identify the restrictions of data sources according to the following characteristics:

1. Attributes that can be specified in the query
2. Attributes that can be included in the predicate of the query

3. Attributes that cannot be specified in the query
4. Attributes that cannot be included in the query predicate.
5. Minimum number of attributes that must be in the query predicate
6. Maximum number of attributes that must be in the query predicate

The strategy to obtain this information is based on the execution of prove-queries. These queries are light queries that belong to a defined set of queries each one of them intended to validate one or more of the restrictions. An example of a prove-query ask for a specific attribute without any predicate, this query contributes to the validation of restriction characteristics 1,3 and 5. For evaluating restriction number 2, MetaExtractor tries to execute a query with one condition in the predicate including the attribute compared to a valid value according to the type of the attribute (e.g. numeric, string). If the query does not return an error, the attribute is included in the set of safe attributes in the predicate. The last restriction is evaluated through the incremental of the number of conditions in the predicate. The attributes used in this case are only the ones that passed the restriction 2. Once a data source is registered the defined set of prove-queries is executed over it.

5 Prototype and Functionality Evaluation

In order to evaluate the behavior of MetaExtractor a prototype has been constructed and used to evaluate its extractors and query capabilities. This section presents the main results obtained during this evaluation. Section 5.1 presents the characteristics of the prototype. Section 5.2 details the experimental results.

5.1 Prototype

In order to evaluate the behavior of MetaExtractor we developed the components presented in Figure 4. Components were written in Java. the Global schema and metadata repository are stored in Virtuoso [26]. The communication and interaction of agents uses the multiagent framework BESA[27]. Queries are accepted in SPARQL [28]. The content metadata extractor agent uses the natural language processing API provided by LingPipe [29] and the Weka environment.

Structural Metadata and Global Schema. One of the main issues in mediation systems is how to control the heterogeneity of data sources that contribute to the system. Approaches that allow each data source to manage its own schema have demonstrated to scale well, but have important drawbacks in the expressiveness of domain concepts [30]. Because of this, MetaExtractor follows a global schema approach that allows to relate the structure of each one of the data sources to the schema of the domain that allows to create VDOs. This decision was made considering that mediation systems are used to create virtual environments of data sharing around a domain. The global schema defines this

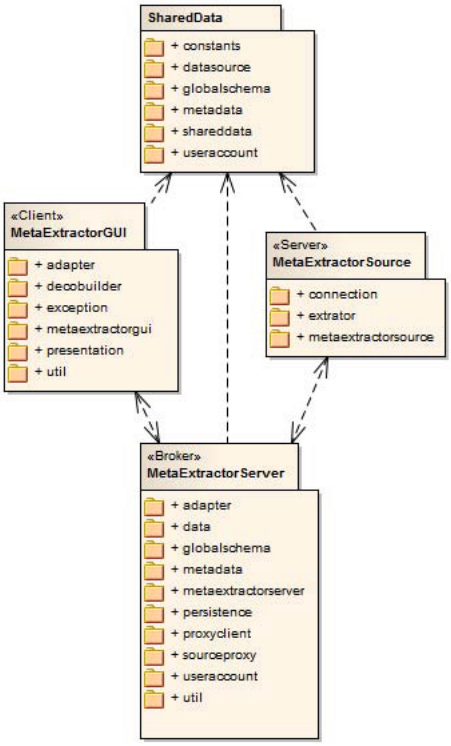


Fig. 4. MetaExtractor Prototype Components

domain, the relationship between a data source and the VDOs global schema represents the contribution of this data source to the domain.

The Global Schema in MetaExtractor is created as an ontology in OWL [31]. It can be created by importing it using a *.owl* extension or automatically based on a seminal data source. In the latter case, the tables of the data source are interpreted as *Classes*, its attributes as *Data Type Properties* and the foreign keys as *Object Type Properties*. Although the extraction from a structured data source does not allow to include all the elements of a OWL ontology, they can be added manually after its creation.

In order to create the relationship between data sources and global schema, MetaExtractor provides a method that identifies synonyms between the names of tables and attributes with the names of classes and properties of the global schema. The execution of this method is not mandatory, but it facilitates the process of data source registration in MetaExtractor.

Metadata Repository Representation. Similarly to the Global Schema, Metadata elements in MetaExtractor are described using the syntax and semantics of an ontology. This ontology includes three main classes: *Data Source,*

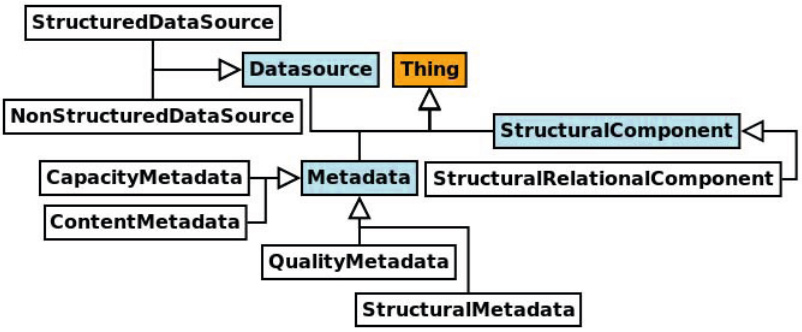


Fig. 5. Metadata Ontology

Metadata, Structural Component. The class *Data Source* has two specializations: *Strucured* and *Non-Structured*, the former one includes a restriction that specifies that an individual of this class must have at least one structural component related. *Metadata* class has four specializations, *Structure*, *Quality*, *Capacity* and *Content*. Each one of them has specializations and related properties. Finally, The Structural Component class and its subclasses define the specialization and elements of structured data sources like tables, attributes, etc. Figure 5 illustrates an excerpt of this ontology.

Metadata pieces are stored as triples *(Subject, Predicate, Object)*, respecting the semantics and syntaxis of metadata ontology and global schema ontology. In these triples *Subject* is an individual of the classes defined in the ontology, *Predicate* is a property and *Object* is the value associated to the *Subject-Predicate*.

When a data source is registered in MetaExtractor the first performed action is to create a triple that specifies that the name of the data source is a *Data Source*. Then, MetaExtractor proceeds creating the triple that describes the data source. The following expression is an example on how to create a triple:

```
Triple triple = new Triple(
Node.createURI(systemOntologyBaseName + ‘‘\#’’ + tableName),
Node.createURI(systemOntologyBaseName + ‘‘\#hasAttribute’’),
Node.createLiteral(systemOntologyBaseName + ‘‘\#’’ + attribute))
```

5.2 Experimental Results

Eventhough MetaExtractor was created for being used in any context, the motivational context was the integration of medical data sources containing electronic health records in a country. Tests with MetaExtractor were executed in this context. The environment involves 100 simulated data sources containing EHRs. Data sources were created in three different DBMS: PostgreSQL, MySQL and SQL Server 2012. The records contained in each data source were extracted from a database from a general hospital; however, for testing purposes, the database

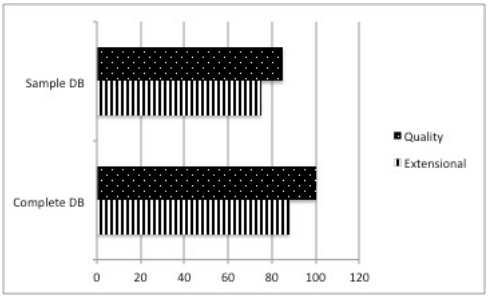


Fig. 6. Precision Quality and Content Metadata with Samples

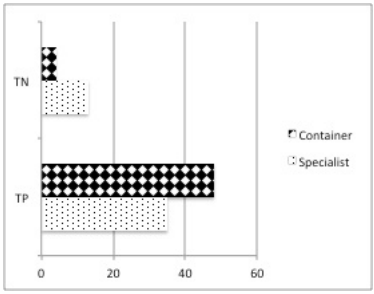


Fig. 7. True Positives and True Negatives Specialists and Containers

was divided into subsets in order to have different types of specializations (e.g. data sources specialized on obstetrics and gynecology) or general data sources. Categories were trained based on the SNOMED terms.

Tests involve the extraction of structure, content, quality and capability metadata as well as the prioritization using the extracted metadata. Due to space limitations we are going to focus in content metadata extraction tests. Figure 6 shows the results comparing the precision of metadata when the extraction of content and quality metadata used samples instead of the complete databases. Although the extraction of samples reduces the precision of the extracted metadata, the tests allows us to conclude that the precision of metadata is stable and can be used to extrapolate the quality and content of the complete data source.

Figure 7 illustrates the differences on True Positives (TP) and False Positives (FP) taking into account the Container and Specialist roles. The main conclusion of these tests is that the rate of TP and FP confirms the reliability of the system.

6 Conclusions and Future Work

This paper presents MetaExtractor a system that extracts, stores and maintains metadata. MetaExtractor is able to extract content, quality, capability and structure metadata from structured data sources that contain high volume of narrative texts. To the best of our knowledge, this is the first attempt to support formally the extraction of this type of metadata in mediation systems.

MetaExtractor architecture is based on the interaction of software agents in charge of extracting different types of metadata from distributed data sources. It stores metadata in a repository that follows the semantics and syntax of an ontology that allows to obtain precise metadata given a user query. MetaExtractor was designed to be used in architectures where the proportion of distribution with respect to the structure of an object (VDO) is higher than the distribution of its instances, making the metadata useful to reduce the number of sources required for a query. It improves the evaluation of queries that involve predicates with high selectivity and guarantees the dynamic adaptability of the system.

MetaExtractor is being implemented as part of a large scale mediation system. The current prototype is a proof-of-concept. Performance evaluation will be tested in the next stage of development. Short term work involves the analysis of redundancy between data sources to avoid the use of data sources that may provide the same set of instances.

Acknowledgements. This work was supported by the project *"Extracción semi-automática de metadatos de fuentes de datos estructuradas: Una aproximación basada en agentes y minería de datos"* funded by Banco Santander S.A. and Pontificia Universidad Javeriana.

References

1. Doan, A., Halevy, A.Y.: Efficiently ordering query plans for data integration. In: ICDE 2002, p. 393. IEEE Computer Society, Washington, DC (2002)
2. Akbarinia, R., Martins, V.: Data management in the appa system. Journal of Grid Computing (2007)
3. Bleiholder, J., Khuller, S., Naumann, F., Raschid, L., Wu, Y.: Query planning in the presence of overlapping sources. In: EDBT, pp. 811–828 (2006)
4. Hayek, R., Raschia, G., Valduriez, P., Mouaddib, N.: Summary management in p2p systems. In: EDBT, pp. 16–25 (2008)
5. Sarma, A.D., Dong, X.L., Halevy, A.: Data integration with dependent sources. In: Proceedings of the 14th International Conference on Extending Database Technology, EDBT/ICDT 2011, pp. 401–412. ACM, New York (2011)
6. Wiederhold, G.: Mediators in the architecture of future information systems. Computer 25, 38–49 (1992)
7. Roth, M., Schwarz, P.: A wrapper architecture for legacy data sources. In: VLDB 1997, pp. 266–275. Morgan Kaufmann (1997)
8. Tomasic, A., Raschid, L., Valduriez, P.: Scaling access to heterogeneous data sources with DISCO. Knowledge and Data Engineering 10, 808–823 (1998)
9. Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J.D., Vassalos, V., Widom, J.: The tsimmis approach to mediation: Data models and languages. Journal of Intelligent Information Systems 8, 117–132 (1997)
10. Kossmann, D.: The state of the art in distributed query processing. ACM Comput. Surv. 32, 422–469 (2000)
11. Levy, A.Y., Rajaraman, A., Ordille, J.J.: Querying heterogeneous information sources using source descriptions. In: VLDB, pp. 251–262 (1996)
12. Pottinger, R., Halevy, A.Y.: Minicon: A scalable algorithm for answering queries using views. VLDB Journal. 10, 182–198 (2001)
13. Khatri, H., Fan, J., Chen, Y., Kambhampati, S.: Qpiad: Query processing over incomplete autonomous databases. In: ICDE, pp. 1430–1432 (2007)
14. Naumann, F., Freytag, J.-C., Leser, U.: Completeness of integrated information sources. Inf. Syst. 29, 583–615 (2004)
15. Huebsch, R., Hellerstein, J.M., Lanham, N., Loo, B.T., Shenker, S., Stoica, I.: Querying the internet with pier. In: VLDB, pp. 321–332 (2003)
16. Villamil, M.D.P., Roncancio, C., Labbe, C.: Pins: Peer-to-peer interrogation and indexing system. In: IDEAS 2004: Proceedings of the International Database Engineering and Applications Symposium, pp. 236–245. IEEE Computer Society, Washington, DC (2004)

17. Tatarinov, I., Ives, Z., Madhavan, J., Halevy, A., Suciu, D., Dalvi, N., Dong, X.L., Kadiyska, Y., Miklau, G., Mork, P.: The piazza peer data management project. SIGMOD Rec. 32, 47–52 (2003)
18. Ooi, B.C., Tan, K.L., Zhou, A., Goh, C.H., Li, Y., Liau, C.Y., Ling, B., Ng, W.S., Shu, Y., Wang, X., Zhang, M.: Peerdb: peering into personal databases. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, SIGMOD 2003, pp. 659–659. ACM, New York (2003)
19. Rousset, M.-C., Adjiman, P., Chatalic, P., Goasdoué, F., Simon, L.: SomeWhere: A scalable peer-to-peer infrastructure for querying distributed ontologies. In: Meersman, R., Tari, Z. (eds.) OTM 2006, Part I. LNCS, vol. 4275, pp. 698–703. Springer, Heidelberg (2006)
20. Greenberg, J.: Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. Journal of Internet Cataloging 6, 59–82 (2004)
21. Margaritopoulos, M., Margaritopoulos, T., Kotini, I., Manitsaris, A.: Automatic metadata generation by utilising pre-existing metadata of related resources. Int. J. Metadata Semant. Ontologies 3, 292–304 (2008)
22. Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., Melton, G.B.: Using discordance to improve classification in narrative clinical databases: An application to community-acquired pneumonia. Comput. Biol. Med. 37, 296–304 (2007)
23. Kowalski, G.: Information Retrieval Systems: Theory and Implementation, 1st edn. Kluwer Academic Publishers, Norwell (1997)
24. Pomares-Quimbaya, A., Roncancio, C., Cung, V.D., Villamil, M.D.P.: Improving source selection in large scale mediation systems through combinatorial optimization techniques. T. Large-Scale Data- and Knowledge-Centered Systems 3, 138–166 (2011)
25. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th edn. John Wiley and Sons (2011)
26. Software, O.: Openlink software virtuoso open-source (vos) versión 1.7.2.1 (2011)
27. González, E., Bustacara, C.J., Avila, J.A.: Besa: Arquitectura para construcción de sistemas multiagentes. In: CLEI - Conferencia Latinoamericana de Estudios en Informática Ponencia (2003)
28. Eric Prud, A.S.: Sparql query language for rdf (2007), `http://www.w3.org/tr/rdf-sparql-query/`
29. Carpenter, B., Baldwin, B.: Text Analysis with Ling Pipe 4. Ling Pipe Publishing (2011)
30. Ooi, B.C., Tan, K.L., Zhou, A., Goh, C.H., Li, Y., Liau, C.Y., Ling, B., Ng, W.S., Shu, Y., Wang, X., Zhang, M.: Peerdb: Peering into personal databases. In: SIGMOD Conference, p. 659 (2003)
31. W3C OWL Working Group, OWL 2 Web Ontology Language: Document Overview. W3C Recommendation (October 27, 2009), `http://www.w3.org/TR/owl2-overview/`
32. International Health Terminology Standards Development Organisation. SNOMED Clinical Terms (2013), `http://www.ihtsdo.org/snomed-ct/`

Proxy Service for Multi-tenant Database Access

Haitham Yaish[1,2], Madhu Goyal[1,2], and George Feuerlicht[2,3]

[1] Centre for Quantum Computation & Intelligent Systems
[2] Faculty of Engineering and Information Technology
University of Technology, Sydney
P.O. Box 123, Broadway NSW 2007, Australia
[3] Faculty of Information Technology,
University of Economics, Prague, Czech Republic
haitham.yaish@student.uts.edu.au, madhu@it.uts.edu.au,
george.feuerlicht@uts.edu.au

Abstract. The database of multi-tenant Software as a Service (SaaS) applications has challenges in designing and developing a relational database for multi-tenant applications. In addition, combining relational tables and virtual relational tables to make them work together and act as one database for each single tenant is a hard and complex problem to solve. Based on our multi-tenant Elastic Extension Tables (EET), we are proposing in this paper a multi-tenant database proxy service to combine multi-tenant relational tables and virtual relational tables, to make them act as one database for each single tenant. This combined database is suitable to run with multi-tenant SaaS single instance applications, which allow tenants designing their database and automatically configuring its behavior during application runtime execution. In addition, these applications allow retrieving tenants data by simply calling functions from this service which spare tenants from spending money and efforts on writing SQL queries and backend data management codes, and instead allowing them to focus on their business and to create their web, mobile, and desktop applications. The feasibility and effectiveness of the proposed service are verified by using experimental data on some of this service functions.

Keywords: Software as a Service, SaaS, Multi-tenancy, Multi-tenant Database, Relational Tables, Virtual Relational Tables, Elastic Extension Tables.

1 Introduction

Configuration is the main characteristic of multi-tenant applications that allow SaaS vendors running a single instance application, which provides a means of configuration for multi-tenant applications. This characteristic requires a multi-tenant aware design with a single codebase and metadata capability. Multi-tenant aware application allows each tenant to design different parts of the application, and automatically adjust and configure its behavior during runtime execution without redeploy the application [3]. Multi-tenant data has two types: shared data, and tenant's isolated data. By combining these data together tenants can have a complete data which suits their business needs [5][11].

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 100–117, 2013.

There are various models of multi-tenant database schema designs and techniques which have been studied and implemented to overcome multi-tenant database challenges [14]. Nevertheless, these techniques are still not overcoming multi-tenant database challenges [1]. NoSQL stands for Not Only Structured Query Language, is a non-relational database management system. This technique avoids join operations, filtering on multiple properties, and filtering of data based on the results of a subqueris. Therefore, the efficiency of NoSQL simple query is very high, but this is not the case for complex queries [4][10][6]. Salesforce.com [13], the pioneer of SaaS Customer Relationship Management (CRM) applications has developed a storage model to manage its virtual database structure by using a set of metadata, universal data table, and pivot tables. Also, it provides a special object-oriented procedural programming language called Apex, and two special query languages: Sforce Object Query Language (SOQL) and Sforce Object Search Language (SOSL) to configure, control, and query the data from Salesforce.com storage model [9].

We have proposed a novel multi-tenant database schema design to create and configure multi-tenant applications, by introducing an Elastic Extension Tables (EET) which consists of Common Tenant Tables (CTT) and Virtual Extension Tables (VET) . The database design of EET technique is shown in the Appendix. This technique enables tenants creating and configuring their own virtual database schema including: the required number of tables and columns, the virtual database relationships for any of CTTs or VETs, and the suitable data types and constraints for a table columns during multi-tenant application run-time execution [14]. In this paper, we are proposing a multi-tenant database proxy service called Elastic Extension Tables Proxy Service (EETPS) to combine, generate, and execute tenants' queries by using a codebase solution that converts multi-tenant queries into a normal database queries.

Our EETPS provides the following new advancements:

- Allowing tenants to choose from three database models. First, multi-tenant relational database. Second, combined multi-tenant relational database and virtual relational database. Third, virtual relational database.
- Avoiding tenants from spending money and efforts on writing SQL queries, learning special programing languages, and writing backend data management codes by simply calling functions from our service which retrieves simple and complex queries including join operations, filtering on multiple properties, and filtering of data based on subqueries results.

In our paper, we explored two sample algorithms for two functions of our service, and we carried out four types of experiments to verify the practicability of our service.

The rest of the paper is organized as follows: section 2 reviews related work. Section 3 describes Elastic Extension Tables Proxy Service, section 4 describes two sample algorithms of the Elastic Extension Tables Proxy Service, section 5 gives our experimental results and section 6 concludes this paper and descries the future work.

2 Related Work

There are various models of multi-tenant database schema designs and techniques which have been studied and implemented to overcome multi-tenant database

challenges like Private Tables, Extension Tables, Universal Table, Pivot Tables, Chunk Table, Chunk Folding, and XML [1][2][7][8][14]. Nevertheless, these techniques are still not overcoming multi-tenant database challenges [1]. Salesforce.com, the pioneer of SaaS CRM applications has designed and developed a storage model to manage its virtual database structure by using a set of metadata, universal data table, and pivot tables which get converted to objects that the Universal Data Dictionary (UDD) keeps track of them, their fields and relationships, and other object definition characteristics. Also, it provides a special object-oriented procedural programming language called Apex which does the following. First, declare program variables, constants, and execute traditional flow control statements. Second, declare data manipulation operations. Third, declare the transaction control operations. Then Salesforce.com compiles Apex code and stores it as metadata in the UDD [13]. In addition, it has its own Query Languages, first, Sforce Object Query Language (SOQL), which retrieve data from one object at a time. Second, Sforce Object Search Language (SOSL), which retrieve data from multiple objects simultaneously [9] [13]. NOSQL is a non-relational database management system which designed to handle storing and retrieving large quantities of data without defining relationships. It has been used by cloud services like MongoDB, Cassandra, CouchDB, Google App Engine Datastore, and others. This technique avoids join operations, filtering on multiple properties, and filtering of data based on subqueries results. Therefore the efficiency of its simple query is very high, but this is not the case for complex queries. Moreover, unless configuring NoSQL consistency models in protective modes of operation, NoSQL will not assure the data consistency and it might sacrifice data performance and scalability [4][10]. Indrawan-Santiago [13] states that NoSQL should be seen as a complimentary solution to relational databases in providing enhanced data management capability, not as a replacement to them.

3 Elastic Extension Tables Proxy Service

In this paper, we are proposing a multi-tenant database proxy service to combine, generate, and execute tenants' queries by using a codebase solution which converts multi-tenant queries into normal database queries. This service has two objectives, first, to enable tenants' applications retrieve tuples from CTTs, retrieve combined tuples from two or more tables of CTTs and VETs, or retrieve tuples from VETs. Second, to spare tenants from spending money and efforts on writing SQL queries and backend data management codes by simply calling functions from this service, which retrieves simple and complex queries including join operations, union operations, filtering on multiple properties, and filtering of data based on subqueries results.

This service gives tenants the opportunity of satisfying their different business needs and requirements by choosing from any of the following three database models which are also shown in Fig.1.

- Multi-tenant relational database: This database model eligible tenants using a ready relational database structure for a particular business domain database without any

need of extending on the existing database structure, and this business domain database can be shared between multiple tenants and differentiate between them by using a Tenant ID. This model can be applied to any business domain database like: CRM, Accounting, Human Resource (HR), or any other business domains.

- Combined multi-tenant relational database and virtual relational database: This database model eligible tenants using a ready relational database structure of a particular business domain with the ability of extending on this relational database by adding more virtual database tables, and combine these tables with the existing database structure by creating virtual relationships between them.
- Multi-tenant virtual relational database: This database model eligible tenants using their own configurable database through creating their virtual database structures from the scratch, by creating virtual database tables, virtual database relationships between the virtual tables, and other database constraints to satisfy their special business requirements for their business domain applications.

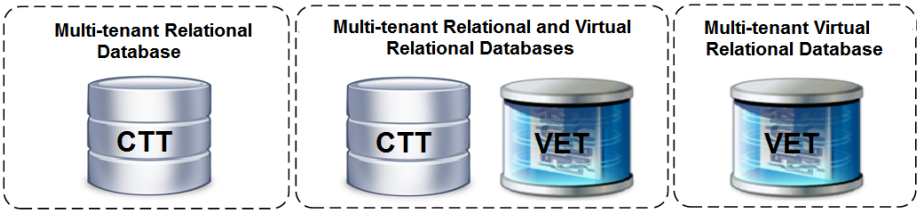


Fig. 1. EETPS database models

The EETPS provides functions which allow tenants building their web, mobile, and desktop applications without the need of writing SQL queries and backend data management codes. Instead, retrieving their data by simply calling these functions, which return a two dimensional array (Object [α] [β]), where α is the number of array rows that represents a number of retrieved tuples, and β is the number of array columns that represents a number of retrieved columns for a particular virtual table. These functions were designed and built to retrieve tenants' data from the following tables:

- One table either a CTT or a VET.
- Two tables which have one-to-one, one-to-many, many-to-many, or self-referencing relationships. These relationships can be between two VETs, two CTTs, or one VET and one CTT.
- Two tables which may have or not have relationships between them, by using different types of joins including: Left Join, Right Join, Inner Join, Outer Join, Left Excluding Join, Right Excluding Join, and Outer Excluding Join. The Join operation can be used between two VETs, two CTTs, or one VET and one CTT.
- Two tables or more which may have or not have relationships between them, by using the union operator that combines the result-set of these tables whether they are CTTs or VETs.

- Two or more tables which have relationships between them, by using filters on multiple tables, or filtering data based on the results of subqueries.

Moreover, the EETPS functions have the capabilities of retrieving data from CTTs or VETs by using the following query options: Logical Operators, Arithmetic operators, Aggregate Functions, Mathematical functions, Using Single or Composite Primary Keys, Specifying Query SELECT clauses, Specifying Query WHERE Clause, Specifying Query Limit, and Retrieving BLOB and CLOB Values.

4 Sample Algorithms of the Elastic Extension Tables Proxy Service

In this section, two sample algorithms will be explored, Single Table Algorithm, and Union Tables Algorithm.

4.1 Single Table Algorithm

This algorithm retrieves tuples from a CTT or a VET. There are three different cases in this algorithm, first, retrieving tuples from a VET by specifying certain primary keys. Second, retrieving tuples from a VET by specifying certain table row IDs which are stored in 'table_row' extension table. Third, retrieving all tuples of a CTT or a VET. In this section we will explore the main algorithm and some of the subsidiary algorithms of the Single Table Algorithm including: the algorithm of the second case that mentioned in this paragraph, and Store Tuples in Array Algorithm. In addition, we will explore an example for each of these algorithms.

Single Table Main Algorithm. This main algorithm is outlined in Program Listing 1. The algorithm determines which of the three cases mentioned above will be applied by checking the passed parameters, and based on these parameters one of a three different query statement will be constructed, and then this query statement will be passed to 'getQuery' algorithm which will return SQL query results from 'table_row', 'table_row_blob', and 'table_row_clob' extension tables and store these results in a set. Then, this set will be passed to Store Tuples in Array Algorithm which will store the results in a two dimensional array, where the number of array rows represents a number of retrieved tuples, and the number of array columns represents a number of retrieved columns for a particular table.

Definition 1 (Single Table Main Algorithm). T denotes a tenant ID, B denotes a table name, λ denotes a set of table row IDs, Ω denotes a set of primary keys, S denotes a string of the SELECT clause parameters, W denotes a string of the WHERE clause, F denotes a first result number of a query limit, M denotes the maximum amount of a query limit which will be retrieved, Q denotes the table type (CTT or VET), I denotes a set of VET indexes, C denotes a set of retrieved tuples from a CTT, V denotes a set of retrieved tuples from a VET, and Φ denotes a two dimensional array that stores the retrieved tuples.

Input.T, B, λ, Ω, S, W, F, M and Q.
Output. Φ.

```
1. if Q = CTT then
2.    C ← retrieve tuples from a CTT by using T, Ω,
      S, W, F, and M to filter the query results
3. else
4.   if Ω ≠ null then
5.      V ← retrieve tuples from a VET by using T,
        Ω, S, W, F,   and M to filter the query
        results
6.   else if λ ≠ null then
7.  /* This statement calls Table Row Query
Algorithm */
8.      V ← retrieve tuples from a VET by using T,
        λ, S, W, F, and M to filter the query
        results
9.   else
10.     I ← retrieve the indexes of B by using
        table_index extension table
11.  end if
12.  if B has I then
13.       V ← retrieve tuples from a VET by using
          T, I, S, W, F, and M to filter the query
          results
14.   else
15.       V ← retrieve tuples from a VET by using
          T, S, W, F, and M to filter the query
          results
16.   end if
17. end if
18. /* This statement calls Store Tuples in Array
Algorithm */
19. store C or V in Φ
20. Return Φ
```

[1] The program listings of Single Table Algorithm.

Table Row Query Algorithm. This subsidiary query algorithm is used to retrieve tuples for a tenant from a VET. The database query which is used in this algorithm uses UNION operator keyword to combine the result-set of three SELECT statements for three tables: table_row, table_row_blob, and table_row_clob if the VET only contains BLOB and/or CLOB, however if the VET does not contain BLOB and CLOB then the UNION operator will not be used in the query.

Definition 2 (Table Row Query Algorithm). T denotes a tenant ID, B denotes a table name, λ denotes a set of table row IDs, S denotes a string of the SELECT clause parameters, W denotes a string of the WHERE clause, F denotes a first result number of the query limit, M denotes the maximum amount of the query limit which will be retrieved, Q denotes the table type (CTT or VET), and θ denotes a string of the select statement.

Input. T, B, λ, Ω, S, W, F, M and Q.
Output. θ.

```
1. θ = SELECT tr.table_column_name, tr.value,
tr.table_row_id, tr.serial_id FROM table_row tr
WHERE tr.tenant_id = T AND tr.db_table_id = B AND
tr.table_row_id IN (λ) AND table_column_id in (S)
AND W
UNION
SELECT trb.table_column_name, trb.value,
trb.table_row_blob_id,trb.serial_id FROM
table_row_blob trb WHERE trb.tenant_id = T AND
trb.db_table_id = B AND trb.table_row_blob_id IN
(λ)
UNION
SELECT  trc.table_column_name, trc.value,
trc.table_row_clob_id, trc.serial_id FROM
table_row_clob trc WHERE trc.tenant_id = T AND
trc.db_table_id = B AND trc.table_row_clob_id IN
(λ)
ORDER BY 3, 4 LIMIT M OFFSET F
2. Return θ
```

[2] The program listings of Table Row Query Algorithm.

Store Tuples in Array Algorithm. This subsidiary algorithm is used to store the retrieved data from a CTT or a VET into a two dimensional array, the number of array rows represents a number of retrieved tuples, and the number of array columns represents a number of retrieved columns for a table. The column names get stored in the first element of this two dimensional array, and the data in these columns get stored in the rest elements of the array.

Definitions 3 (Store Tuples in Array Algorithm). T denotes a tenant ID, B denotes a table name, μ denotes a set of retrieved tuples from a CTT or a VET where each of these tuples is presented as τ and each column of τ is presented as χ , which means τ is a set of χ where $\tau = \{ \chi 1, \chi 2, \ldots, \chi n\}$. δ denotes a set of column names of a CTT or a VET, Φ denotes a two dimensional array to store the retrieved tuples, and τ n (χ m) denotes a value stored in χ m of τ n.

Input. T, B, and μ.
Output. Φ.

```
1. δ ← retrieve the column names of B from
    table_column extension table by using T to
    filter the query results
2. Initialize Φ [size of μ] [size of δ]
3. i ← 0
4. For all column names ∈ δ Do
5.    Φ [0][i] = δi
6.    i ← i + 1
7. end for
8. n ← 0
9. for all τ ∈ μ Do
10.  m ← 0
11.  For all column names ∈ δ Do
12.    Φ [n+1][m] = τ n(χ m)
13.    m ← m + 1
14.  end for
15.  n ← n + 1
16.end for
17.Return Φ
```

[3] The program listings of Store Tuples in Array Algorithm.

Example. This example explores how the Single Table Algorithm retrieves virtual tuples from one VET. There are three cases that this algorithm is handling which mentioned above in this section. In this example we will explore the case where we pass a certain table Row ID to the algorithm. In this example we will pass the following five input parameters:

1. A tenant ID value, which equals 100.
2. A table ID value, which equals 7.
3. A table row ID value, which equals 2.
4. The SELECT clause parameter (S) is empty, this means that the query will retrieve all the columns of the 'store' VET.
5. The WHERE clause (W) is empty, this means that the query is not filtered by the WHERE clause.

Fig. 2 (a) shows the 'store' VET which we will retrieve tuples from. The query in Program Listing 4 is generated by using the Single Table Algorithm to retrieve a virtual tuple from the 'store' VET based on the passed parameters. Fig. 2 (b) shows the result of the virtual tuples that retrieved from table_row extension table by using this query listed in Program Listing 4. This virtual tuple is divided into three physical tuples, each of these physical tuples stores a column name and its value, and all of these tuples are sharing one 'table_row_id' which equals 2. The query in Program

Listing 4 does not contain the UNION part of the query to retrieve BLOB and CLOB values because the 'store' VET structure does not contain any of them.

The two dimensional array that is shown in Fig. 2 (c) illustrates how the previous result which is shown in Fig. 2 (b) is stored in a well structured two dimensional array. The column names are stored in the first row elements, and the first tuple is stored in the second row elements of the array. Compared with the previous results of the tuples that is shown in Fig. 2 (b), this two dimensional array stores the virtual tuple in a structure which is very similar to any physical tuple that is structured in any physical database table, which in return will facilitate accessing virtual tuples from anyVET.

```
SELECT tr.table_column_name, tr.value,
tr.table_row_id, tr.serial_id FROM table_row tr WHERE
tr.tenant_id = 100 AND tr.db_table_id = 7 AND
trb.table_row_id IN (2)
```

[4] The Program Listing of the query generated by using the Single Table Algorithm.

(a)

store_id	name	phone

(b)

table_column_name	value	table_row_id	serial_id
store_id	2	2	1
name	George Street Store	2	2
phone	+61294455331	2	3

(c)

	0	1	2
0	name	phone	store_id
1	George Street Store	+61294455331	2

Fig. 2. The 'store' VET and some tuples retrieved from it and stored in an array

4.2 Union Tables Algorithm

In this section, we will explore the union function, which retrieves a combined result-set of two or more tables whether they are CTTs or VETs, and stores the result-set in an array. In addition, we will explore an example of this algorithm. The input parameters of this algorithm will determine a tenant, a set of CTTs and/or VETs that the union function needs to retrieve data from, SELECT clauses, and WHERE clauses which are required for each table. Program Listing 5 is showing the detailed algorithm. This algorithm will store the retrieved tuples in an array by using the subsidiary algorithm that mentioned in the Program Listings 3.

Definition 4 (Union Tables Algorithm). T denotes a tenant ID, Π denotes a set of CTTs and VET names, where each of these tables has got one or more tuples (Π = { τ 1, τ 2, …, τ m}), each tuple is presented as τ and each column of τ is presented as χ,

which means τ is a set of χ where τ = { χ1, χ2, …, χn}. υ denotes a set of table columns which are related to the set Π and the columns are ordered according to the table orders, W denotes a set of WHERE clauses which are related to the set Π and the columns are ordered according to the table orders of Π, F denotes a first result number of a query limit, M denotes a maximum amount of a query limit which will be retrieved, Q denotes the table type (CTT or VET), C denotes a set of retrieved tuples from CTT, V denotes a set of retrieved tuples from VET, Φ denotes a two dimensional array which stores the retrieved tuples, and τ n (χ m) denotes a value stored in χ m of τ n.

Input. T, Π, υ, W, F, and M.
Output. Φ.

```
1. i ← 0
2. For all tables ∈ Π Do
3.   if Q = CTT then
4.      C ← retrieve τ from a CTT by using υ, W, F,
        and M to  filter the query results
5.   else
6.      V ← retrieve tuples from a VET by using υ,
        W, F, and (M * size of υ) to filter the
        query results
7.   end if
8.   n ← 0
9.   for all τ ∈ Πi Do
10.    m ← 0
11.    For all column names ∈ τ Do
12.       Φ[n+1][m] = τ n(χ m)
13.       m ← m + 1
14.    end for
15.    n ← n + 1
16.  end for
17.  i ← i + 1
18.end for
19.Return Φ
```

[5] The program listings of Union Tables Algorithm.

Example. This example explores how the Union Table Algorithm retrieves tuples from two tables, the first one CTT and the second one VET. In this example we will pass to the algorithm the following six input parameters:

1. A tenant ID value, which equals 1000.
2. A set of table IDs (Π) which equals {product, 17} where 'product' is a CTT that is shown in Fig. 3(a) and the ID 17 is the ID which represents the 'sales_fact' VET that is shown in Fig. 3 (b).

3. A set of table columns (υ), which equals {{shr_product_id, price}, {58,61}}, where this set contains two other sets, the first one contains the columns of 'product' CTT, and the second one contains the IDs of 'sales_fact' VET. ID 58 represents the virtual 'product_id' column and ID 61 represents the virtual 'unit_price' column.
4. The set of WHERE clauses of the tables (W) are empty, because this example has not got any WHERE clauses parameter passed to the function to filter the tables queries.
5. The first number of the query limits (F), which equals 0.
6. The maximum amount of the query limits (M), which equals 1.

After we passed the parameters to the function, the function iterated the set of tables (Π) , the first table in the set was 'product_id' CTT, the function executed the query which is shown in Program Listing 6 to retrieve the tuples of this table, and the results of this query are shown in Fig. 3 (c). The second table in the set was the 'sales_fact' VET with ID equals 17, the function executed the query in Program Listing 7 and 8. The query in Program Listing 7 was used to retrieve the indexes of the 'sales_fact' VET from 'table_index' extension table, and the query in Program Listing 8 was used to retrieve the virtual tuples from 'sales_fact' VET by using the passed parameters and the 'table_row_id' which were retrieved from the query that shown in Program Listing 7. The results of the two queries of Program Listing 7 and 8 are shown in Fig. 3 (d) and (e).

Finally, the output of the queries of the CTT and the VET that mentioned above are stored in two dimensional array as shown in Fig. 3 (f), the two elements [0] [0] and [0] [1] represent the column names, the Union functions shows generic names like column1, and column 2, however the other functions which our service provides show column names of CTT and VET. The two elements [1] [0] and [1] [1] represent the column's values of the CTT, and the two elements [2] [0] and [2] [1] represent the column's values of the VET.

```
SELECT product_id, price FROM product WHERE tenant_id = 1000 LIMIT 1;
```

[6] The program listing of the query which retrieved the tuples of the 'product' CTT.

```
SELECT table_row_id FROM table_index WHERE tenant_id=1000 AND db_table_id=17 AND (table_column_id=61 OR table_column_id=58) LIMIT 1
```

[7] The program listing of the query which retrieved the indexes of the 'sales_fact' VET from 'table_index' extension table.

```
SELECT tr.table_column_id ,tr.value ,tr.table_row_id, tr.serial_id FROM table_row tr WHERE tr.tenant_id =1000 AND tr.db_table_id = 17 AND tr.table_row_id IN (352871) AND tr.table_column_id in (58,61)
ORDER BY 3,4 LIMIT 2 OFFSET 0
```

[8] The program listing of the query which retrieved the tuples of the 'sales_fact' VET.

(a)

product_id	tenant_id	product_bus_id	standard_cost	price	…

(b)

sales_fact_id	tenant_id	product_id	customer_id	sales_person_id	unit_price	…

(c)

product_id	price
100	5714.87

(d)

table_row_id
352871

(e)

table_column_id	value	table_row_id	serial_id
58	100	352871	4
61	15786	352871	7

(f)

	0	1
0	column 1	column 2
1	100	5714.87
2	100	15786

Fig. 3. The 'product' CTT and the 'sales_fact' VET data structures

5 Performance Evaluation

After developing the EETPS, we carried out four types of experiments to verify the practicability of our service. These experiments were classified according to the complexities of the queries which used in these experiments including: simple, simple-to-medium, medium, and complex. The four experiments show comparisons between the response time of retrieving data from CTTs, VETs, or both CTTs and VETs. We have evaluated the response time through accessing the EETPS which converts multi-tenant queries into normal database queries, instead of accessing the database directly.

5.1 Experimental Data Set

The EETPS has designed and developed to serve multi-tenants in one instance application. However, in this paper the aim of the experiments is to evaluate the performance differences between retrieving the data of CTTs, VETs, or both CTTs and VETs together for one tenant. In our experiment settings we used one machine and we ran the following four types of experiments:

- Simple query experiment (Exp. 1): In this experiment we called a function which retrieved data from a CTT by executing Query 1 (Q1), and retrieved the same data from a VET by executing Query 2 (Q2).
- Simple-to-medium query experiment (Exp. 2): In this experiment we called a function which retrieved data from two CTTs by executing Query 3 (Q3), two VETs by executing Query 4 (Q4), and CTT-and-VET by executing Query 5 (Q5). Each of these two tables combination has got one-to-many relationship between them.
- Medium query experiment (Exp. 3): In this experiment we called a function which retrieved data from two tables by using a union operator for two CTTs by executing Query 6 (Q6), for two VETs by executing Query 7 (Q7), and for CTT-and-VET by executing Query 8 (Q8).

- Complex query experiment (Exp. 4): In this experiment we called a function which uses a left join operator that joined two CTTs by executing Query 9 (Q9), two VETs by executing Query 10 (Q10), and CTT-and-VET by Query 11 (Q11).

In these four experiments we ran the test on eleven queries twice, the first test was to retrieve only 1 tuple, and the second test was to retrieve a 100 of tuples by using the same queries. The queries that we ran on CTTs are the same queries we ran on VETs, and CTT-and-VET in order to have accurate comparisons. The structures of these queries are shown in Fig. 4. We recorded the execution time of these queries experiments based on six data sets for all the four types of experiments that we ran. The first data set contained 500 tuples, the second data set contained 5,000 tuples, the third data set contained 10,000 tuples, the fourth data set contained 50,000 tuples, the fifth data set contained 100,000 tuples, and the last data set contained 200,000 tuples. All of these data sets were for one tenant.

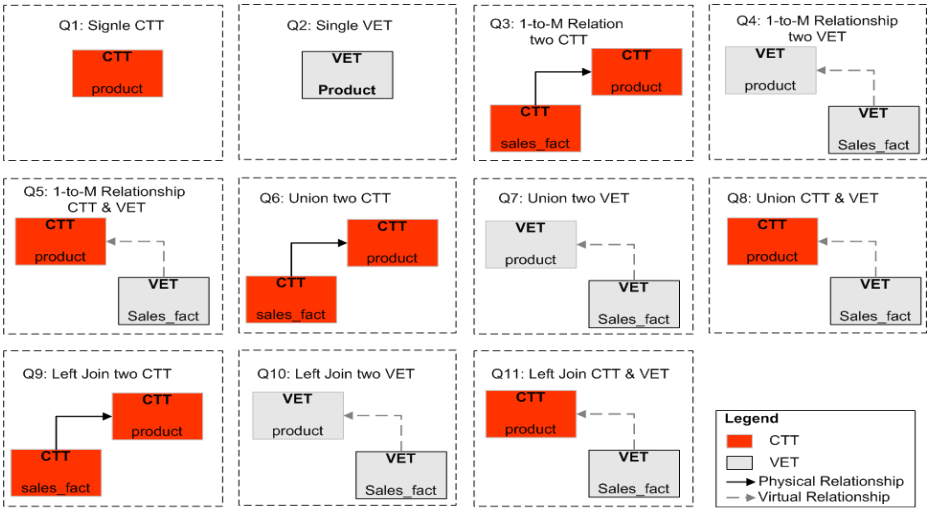


Fig. 4. The structures of the queries executed in our experiments

5.2 Experimental Setup

Our EETPS was implemented in Java 1.6.0, Hibernate 4.0, and Spring 3.1.0. The database is PostgreSQL 8.4 and the application server is Jboss-5.0.0.CR2. Both of database and application server is deployed on the same PC. The operating system is windows 7 Home Premium, CPU is Intel Core i5 2.40GHz, the memory is 8GB, and the hard disk is 500G.

5.3 Experimental Results

In all the experimental diagrams we provided in this section the vertical axes which are the execution time in seconds, and the horizontal axes which are the total number of

tuples that stored in a tenant's tables. Each of the four experiments retrieves 1 tuple and 100 of tuples, and we will show in this section the average execution time of the six data sets of these tuples which are related to CTTs, VETs, and CTTs and VETs, and show the differences between them. These experimental diagrams are shown in Fig 5-12.

We found in our experimental results that the average performance of the CTT and the VET for Exp.1 can be considered the same, and the VETs, the CTTs and the CTT-and-VET for Exp. 2 can be considered the same as well. In addition, we found that the average performance for Exp. 3 for the VETs, and the CTT-and-VET can be considered slightly higher than the CTTs, but the average performance of the VETs is the highest difference between the three types of tables. The average performance difference between the CTTs and the VETs for retrieving 1 tuple is 280 milliseconds, and for retrieving 100 tuples is 396 milliseconds. In the last experimental results Exp. 4 we found that the average performance for the CTT-and- VET can be considered higher than the CTTs by approximately 1.2 seconds, and for the VETs can be considered higher than the CTTs by approximately 1.5 seconds. The details of the experimental results summary are shown in Table 1 and 2.

Table 1. This table shows the experimental results of retrieving 1 tuple in milliseconds

Retrieving 1 Tuple	CTT	VET	CTT-and-VET	Difference Between CTT–and-VET	Difference Between CTT and CTT-and-VET
Exp. 1	**Q 1**	**Q 2**			
	117	161		44	
Exp. 2	**Q 3**	**Q 4**	**Q 5**		
	146	155	149	9	3
Exp. 3	**Q 6**	**Q 7**	**Q 8**		
	231	511	340	280	109
Exp. 4	**Q 9**	**Q 10**	**Q 11**		
	403	1930	1632	1527	1229

Table 2. This table shows the experimental results of retrieving 100 tuples in milliseconds

Retrieving 100 Tuples	CTT	VET	CTT-and-VET	Difference Between CTT–and-VET	Difference Between CTT and CTT-and-VET
Exp. 1	Q 1	Q 2			
	204	206		2	
Exp. 2	Q 3	Q 4	Q 5		
	331	355	343	24	12
Exp. 3	Q 6	Q 7	Q 8		
	245	641	388	396	143
Exp. 4	Q 9	Q 10	Q 11		
	560	2112	1856	1552	1296

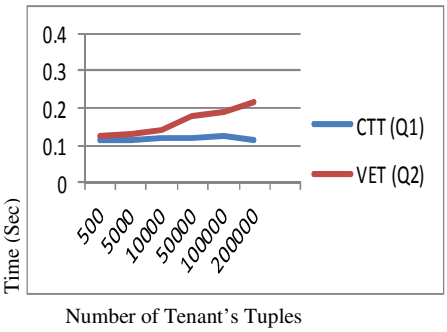


Fig. 5. Single Table 1 Tuple

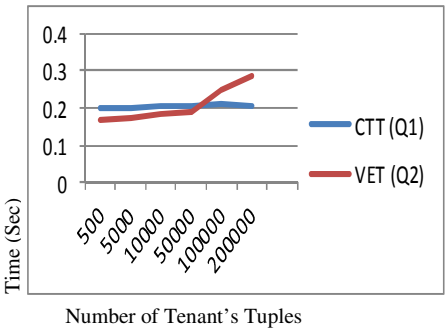


Fig. 6. Single Table 100 Tuples

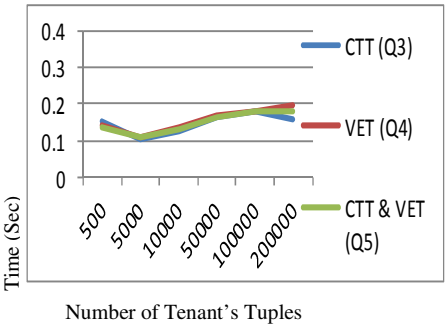


Fig. 7. 1-to-M 1 Tuple

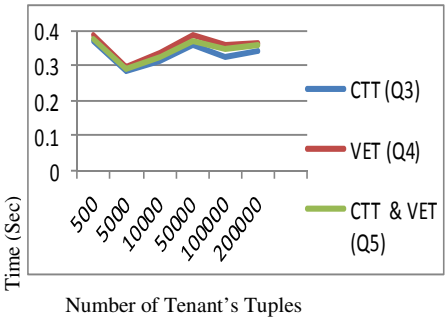


Fig. 8. 1-to-M 100 Tuples

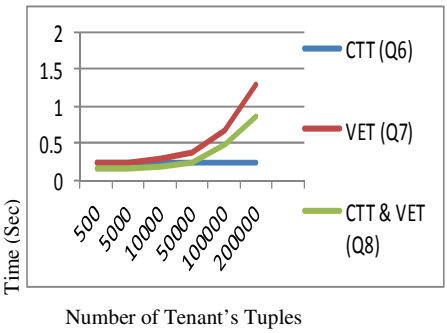


Fig. 9. Union 1 Tuple

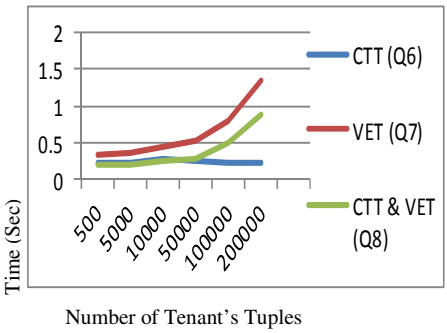


Fig. 10. Union 100 Tuples

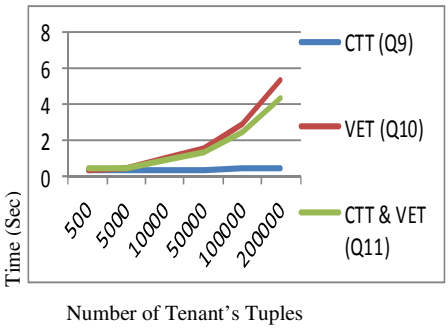


Fig. 11. Left Join 1 Tuple

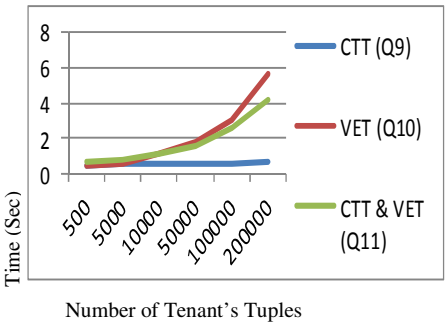


Fig. 12. Left Join 100 Tuples

6 Conclusion

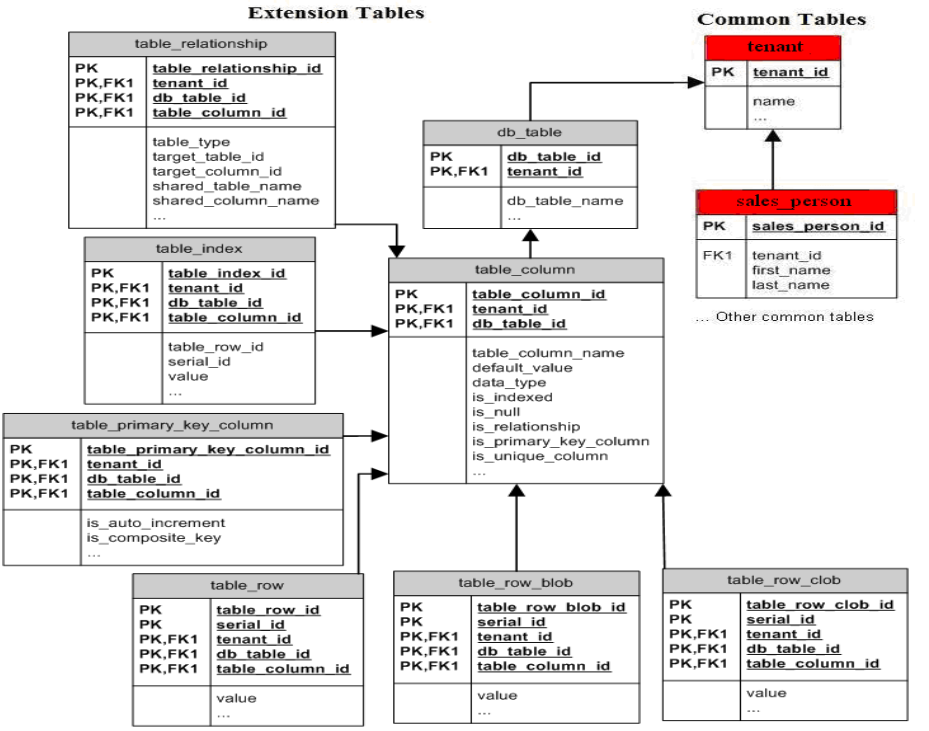
In this paper, we are proposing a multi-tenant proxy service for the EET to combine, generate, and execute tenants' queries by using a codebase solution that converts a multi-tenant query into a normal database query. This service has two objectives, first, allowing tenants to choose from three database models: multi-tenant relational database, combined multi-tenant relational database and virtual relational database, and virtual relational database. Second, sparing tenants from spending money and efforts on writing SQL queries and backend data management codes by calling our service functions which retrieve simple and complex queries including join operations, filtering on multiple properties, and filtering of data based on subqueries results. In our paper, we explored two sample algorithms for two functions, and we carried out four types of experiments to verify the practicability of our service. These experiments were classified according to the complexities of the queries which used in these experiments including: simple, simple-to-medium, medium, and complex. The four experiments show comparisons between the response time of retrieving data from CTTs, VETs, or both CTTs and VETs. In our experimental results we found that the average performance of CTTs, VETs, and CTT- and-VET for the simple queries and the simple-to-medium queries are considered almost the same. Also, we found that the average performance of the medium queries for VETs, and CTT-and-VET is considered slightly higher than CTTs, but VET are the highest between the three types of tables. In the last experimental results of complex query we found that the average performance for CTT-and-VET is considered higher than CTTs by approximately 1.2 seconds, and for VETs is considered higher than CTTs by approximately 1.5 seconds. The cost of complex query is acceptable in favor of obtaining a combined relational database and virtual relational database for multi-tenant applications, which in return these combined databases provide a means of configuration for multi-tenant applications, reduce the Total Cost of Ownership (TCO) on the tenants, and reduce the ongoing operational costs on the service providers.

Our future work will focus on optimizing virtual data retrieval from our EET for simple and complex queries by using a highly-optimized executing query plans and logic, and add more functions to insert, updated, delete tuples from CTT and VET.

References

1. Aulbach, S., Grust, T., Jacobs, D., Kemper, A., Seibold, M.: A Comparison of Flexible Schemas for Software as a Service. In: Proceedings of the 35th SIGMOD International Conference on Management of Data, pp. 881–888. ACM, Rhode Island (2009)
2. Aulbach, S., Grust, T., Jacobs, D., Kemper, A., Rittinger, J.: Multitenant Databases for Software as a Service: Schema Mapping Techniques. In: Proceedings of the 34th SIGMOD International Conference on Management of Data, pp. 1195–1206. ACM, Vancouver (2008)
3. Bezemer, C., Zaidman, A.: Multi-Tenant SaaS Applications: Maintenance Dream or Nightmare? In: Proceedings of the Joint Workshop on Software Evolution and International Workshop on Principles of Software Evolution, pp. 88–92. ACM, Antwerp (2010)
4. Bobrowski, S.: Optimal Multitenant Designs for Cloud Apps. In: 4th International Conference on Cloud Computing, pp. 654–659. IEEE Press, Washington (2012)
5. Domingo, E.J., Nino, J.T., Lemos, A.L., Lemos, M.L., Palacios, R.C., Berbís, J.M.G.: CLOUDIO: A Cloud Computing-Oriented Multi-tenant Architecture for Business Information Systems. In: 3rd International Conference on Cloud Computing, pp. 532–533. IEEE Press, Madrid (2010)
6. Dimovski, D.: Database management as a cloud-based service for small and medium organizations. Master Thesis, Masaryk University Brno (2013)
7. Du, J., Wen, H.Y., Yang, Z.J.: Research on Data Layer Structure of Multi-tenant E-commerce System. In: IEEE 17th International Conference on Industrial Engineering and Engineering Management, Xiamen, pp. 362–365 (2010)
8. Foping, F.S., Dokas, I.M., Feehan, J., Imran, S.: A New Hybrid Schema-sharing Technique for Multitenant Applications. In: Fourth International Conference on Digital Information Management, pp. 1–6. IEEE Press, Michigan (2009)
9. Force.com, `http://www.salesforce.com/us/developer/docs/soql_sosl/salesforce_soql_sosl.pdf`
10. Google Developers, `https://developers.google.com/appengine/docs/python/datastore/overview#Comparison_with_Traditional_Databases`
11. Guoling, L.: Research on Independent SaaS Platform. In: The 2nd IEEE International Conference on Information Management and Engineering, pp. 110–113. IEEE Press, Chengdu (2010)
12. Indrawan-Santiago, M.: Database Research: Are We at a Crossroad? Reflection on NoSQL. In: 15th International Conference on Network-Based Information Systems, pp. 45–51. IEEE Press, Melbourne (2012)
13. Weissman, C.D., Bobrowski, S.: The design of the force.com multitenant internet application development platform. In: Proceedings of the 35th SIGMOD International Conference on Management of Data, pp. 889–896. ACM, Rhode Island (2009)
14. Yaish, H., Goyal, M., Feuerlicht, G.: An Elastic Multi-tenant Database Schema for Software as a Service. In: Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, pp. 737–743. IEEE Press, Sydney (2011)

Appendix: Elastic Extension Tables (EET)



Extracting Correlated Patterns on Multicore Architectures

Alain Casali[1] and Christian Ernst[2]

[1] Laboratoire d'Informatique Fondamentale de Marseille (LIF),
CNRS UMR 6166, Aix Marseille Université
IUT d'Aix en Provence, Avenue Gaston Berger,
13625 Aix en Provence Cedex, France
alain.casali@lif.univ-mrs.fr
[2] Ecole des Mines de St Etienne, CMP - Georges Charpak
880 avenue de Mimet, 13541 Gardanne
ernst@emse.fr

Abstract. In this paper, we present a new approach relevant to the discovery of correlated patterns, based on the use of multicore architectures. Our work rests on a full KDD system and allows one to extract Decision Correlation Rules based on the Chi-squared *criterion* that include a target column from any database. To achieve this objective, we use a levelwise algorithm as well as contingency vectors, an alternate and more powerful representation of contingency tables, in order to prune the search space. The goal is to parallelize the processing associated with the extraction of relevant rules. The parallelization invokes the PPL (Parallel Patterns Library), which allows a simultaneous access to the whole available cores / processors on modern computers. We finally present first results on the reached performance gains.

1 Introduction and Motivation

In the past couple of years, innovations in hardware architecture, like hyper-threading capabilities or multicore processors, have begun to allow parallel computing on inexpensive desktop computers. This is significant in that standard software products will soon be based on concepts of parallel programming implemented on such hardware, and the range of applications will be much broader than that of scientific computing (the main area for parallel computing). Consequently, there is a growing interest in the research field of parallel data mining algorithms, especially in association rules mining (ARM). By exploiting multicore architectures, parallel algorithms may improve both execution time and memory requirement issues, the main objectives of the data mining field.

Independently of this framework, we soon developed a KDD system based on the discovery of Decision Correlation Rules with large and specialized databases (Casali and Ernst, 1). These rules are functional in semiconductor fabrication capabilities. The goal is to discover the parameters that have the most impact on a specific parameter, the yield of a given product ... Decision Correlation

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 118–133, 2013.

Rules (DCRs) are similar to Association Rules. But, as it will be shown, there are huge technical differences, and the overall computation times of DCRs are in the end much better. Furthermore, after implementing DCRs using "conventional sequential algorithms", we decided to adapt this approach to multicore implementation possibilities.

This paper is organized as follows. In Section 2, we recall current features of multicore programming. Section 3 is dedicated to related work: we present (*i*) an overview of Association Rules Mining over a multicore architecture and (*ii*) what Decisional Correlation Rules are. Section 4 describes the concepts used for multicore decision rules mining and our algorithm. In Section 5 we show the results of first experiments. Finally, in the final Section, we summarize our contribution and outline some research perspectives.

2 Recent Advances in Multicore Programmation

Multicore processing is not a new concept, however only recently has the technology has become mainstream with Intel or AMD introducing commercially available multicore chips in 2008. At that date, no software environment able to take advantage simultaneously of the different existing processors had been proposed, let alone produced for the commercial market.

The situation radically changed in 2010. We present in this section these new opportunities, showing the covered aspects in the C++ language used in our developments. We first introduce what the approaches towards parallelization were until only a few years ago. We then present quickly the Lambda Calculus, a new programming paradigm integrated into the last C++ norm and essential to understand multicore prorammation. And we finally expose the PPL environment, which allows effective multicore programmation in a simple way.

2.1 Parallelization Issues on Modern Desktop Computers

For about twenty years, parallelizing tasks on personal computers consisted essentially in the development of multithreaded code, by adding or not higher abstraction levels on the base threaded layers. This often required complex coordination of threads, and introduced difficulties in finding bugs due to the interweaving of processing of data shared between the threads. Although threaded applications added limited amounts of performance on single-processor machines, the extra overhead of development was difficult to justify on the same machines.

But given the increasing emphasis on multicore chip architectures, developers unable to design software to fully exploit the resources provided by multiple cores will now quickly reach performance ceilings.

Multicore processing has thus affected the ability of actual computational software development. Many modern languages do not support multicore functionality. This requires the use of specialized libraries to access code written in languages such as C. There are different conceptual models to deal with the problem, for example using a coordination language and program building blocks

(programming libraries and/or higher order functions). Each block may have a different native implementation for each processor type. Users simply program using these abstractions, and an intelligent compiler then chooses the best implementation based on the context (Darlington, 2).

Many parallel programming models have been recently proposed (Cilk++, OpenMP, OpenHMPP, ...) for use on multicore platforms. Intel introduced a new abstraction for C++ parallelism called TBB. Former and more organization oriented research efforts include the Codeplay Sieve System, Cray's Chapel, Sun's Fortress, and IBM's X10. A comparison of some OpenXXX approaches can be found in (Hamidouche and al., 3).

But the majority of these models roughly bases on the intelligent transformation of general code into multithreaded code. A new, even if simple idea, has been proposed by the OpenMP consortium in FORTRAN in 1997, based on the fact that looping functions are the key area where splitting parts of a loop across all available hardware resources may increase application performance. The Open Multiprocessing Architecture Review Board (or OpenMP ARB) became an API that supports shared memory multiprocessing programming in C++. It consists of a set of compiler directives, library routines, and environment variables that influence run-time behavior. In order to schedule a loop across multiple threads, the OpenMP `pragma` directives were introduced in 2005 to explicitly relay to the compiler more about the transformations and optimizations that should take place. The example chosen to illustrate our purposes is the computation of the sum of an integer vector:

```
vector<int> v = ...;
int sum = 0;
#pragma omp for
for (int i = 0; i < v.size(); i++) {
   #pragma omp atomic
   sum += v[i];
}
```

The first directive requests that the *for* loop should be executed on multiple threads, while the second is used to prevent multiple simultaneous writes to the *sum* variable ...

This approach has then be adapted to multicore programmation, as illustrated hereafter. Let us previously underline that parallelism also has its limits. It is widely admitted that applications that may benefit from using more than one processor necessitate (*i*) operations that require substantial amount of processor time, measured in seconds (or larger denominations) rather than milliseconds and (*ii*) one or more loops of some kind, or operations that can be divided into discrete but significant units of calculation that can be executed independently of one another.

2.2 A New Programming Paradigm: The Lambda Calculus

The example above without parallelization issues can be re-written as follows:

```
vector<int> v = ...;
int sum = 0;
for (int i = 0; i < v.size(); i++){
   sum += v[i];
}
```

A declarative looping technique generalized the syntax for STL containers: the *for_each* function has been introduced in a first C++ 0x draft, so the code can be re-written in:

```
vector<int> v = ...;
int sum = for_each(v.begin(), v.end(), Adder());
```

The third argument is a functor: a class overloading the () operator. In our example, the functor has to perform the addition. This approach did not result in much success for many reasons, the two main of which include: (*i*) developers must implement the functor, but defining a whole class in order to finally execute a single instruction is a bit verbose; (*ii*) using a named function is sub-optimal because confusion may arise with the "C" pointer function syntax.

Recognizing these shortcomings, the C++ draft provided a simplified syntax based on Lambda functions. The associated formalism is a computational model invented by Alonzo Church in the 1930s, which directly inspired both the syntax and the semantics of most functional programming languages (Mitchell, 4). The basic concept of the λ calculus is the "expression", recursively defined as follows:

<expression>:= <name>— <function>— <application>
<function>:= λ <name>.<expression>
<application>:= <expression><expression>

The λ calculus in its most basic form has two operations: (*i*) Abstractions, which correspond to anonymous functions, and (*ii*) Applications, which exist to apply the function. Abstraction is performed using the λ operator, giving the lambda calculus its name. Anonymous functions are often called "lambdas", "lambda functions" or "lambda expressions": these remove all the need for the scaffolding code, and allow a predicate function to be defined in-line in another statement. The current example can thus be re-written as follows:

```
vector<int> v = ...;
int sum = 0;
for_each(v.begin(), v.end(),[&sum](int x) {sum += x;}
);
```

The syntax of a lambda function is reasonably straight-forward, of the form:

```
[lambda-capture] (parameter-list) {->} return-type {statement-list}
```

In the example, the first element of the lambda in the square brackets is called the capture specification: it relays to the compiler that a lambda function is being created and that the local variable *sum* is being captured by reference. The final part is the function body.

Therefore, lambdas behave like function objects, except for we cannot access the class that is generated to implement a lambda in any way other than using the lambda. Consequently, any function that accepts functors as arguments will accept lambdas, but any function only accepting function pointers will not.

These and much more features of lambda functions have been included in the C++11 language norm, allowing a more declarative programming style, taking advantage of (STL) algorithms in a much simpler and cleaner form. Lambda functions allow the inline definition of a function body in the code section in which it is to be logically used. As well as providing strong hints to the compiler about potential real time optimizations, lambda functions make discerning the intent about what a section of code is doing much easier.

2.3 Multicore Programmation Using the PPL

In a different way than the OpenMP consortium, Microsoft developed its own "parallel" approach in the 2000s within an internal work group called Parallel Extensions. Since 2010, and the relevant versions of the .NET framework and Visual Studio, Microsoft enhanced support for parallel programming by providing a runtime tool and a class library among other utilities. The library is composed of two parts: Parallel LINQ (PLINQ), a concurrent query execution engine, and Task Parallel Library (TPL), a task parallelism component of the .NET framework. This component hides entirely the multi-threading activity on the cores: the job of spawning and terminating threads, as well as scaling the number of threads according to the number of available cores, is done by the library itself. The main concept is here a Task, which can be executed independently.

The Parallel Patterns Library (PPL) is the corresponding available tool in the Visual C++ environment, and is defined within the Concurrency namespace. The PPL operates on small units of work (Tasks), each of them being defined by a λ calculus expression. The PPL defines almost three kinds of facilities for parallel processing: (*i*) templates for algorithms for parallel operations, (*ii*) class templates for managing shared resources, and (*iii*) class templates for managing and grouping parallel tasks.

The library provides essentially three algorithms defined as templates for initiating parallel execution on multiple cores:

- The *parallel_for* algorithm is the equivalent of a *for* loop that executes loop iteration in parallel,
- The *parallel_for_each* algorithm executes repeated operations on a STL container in parallel,
- The *parallel_invoke* algorithm executes a set of two or more independent Tasks in parallel, in the sense of different *programs* within the same runtime environment.

Our current example re-written using parallel capabilities:

```
vector<int> v = ...;
int sum = 0;
parallel_for_each(v.begin(), v.end(), [&sum](int x)
 {
   // make sure <sum> is shared between the "cores"...
   sum += x;
 }
);
```

Let us underline again that if the computation on each iteration is very short, there will be inevitably important overhead in allocating the task to a core on each iteration. Which may severely erode any reduction in execution times. This will also be the case if the overall loop integrates important shared resources management, as will be shown in Section 4. This means that upgrading from sequential to parallel computing must be done very carefully.

3 Related Work

Due to the variety of the algorithms (and their specific internal data structures) there does not exist any general model allowing parallel ARM computation. Main techniques are described in Section 3.1. They just are actual models optimized for the usage of multicore architecture. Developers have to write their own thread managers (Herlihy and Shavit, 5). Section 3.2 presents the main results about what Decision Correlation Rules are and how can we compute them using a single processor.

3.1 Association Rules Mining Using Multicore Support

Current research can be divided into three main categories: (*i*) adaptation of the A-Priori algorithm, (*ii*) vertical mining, and (*iii*) pattern-growth method.

A-Priori Based Algorithms: Most of the parallel ARM algorithms are based on parallelization of A-Priori (Agrawal and Srikant, 6) that iteratively generates and tests candidate itemsets from length 1 to k until no more frequent itemsets are found. These algorithms can be categorized into *Count Distribution*, *Data Distribution* and *Candidate Distribution* methods (Agrawal and Shafer, 7), (Han and Kamber, 8). The *Count Distribution* method partitions the database into horizontal partitions, that are independently scanned, in order to obtain the local counts of all candidate on each process. At the end of each iteration, the local counts are summed up a into the global counts so that frequent itemsets can be found. The *Data Distribution* method utilizes the main memory of parallel machines by partitioning both the database and the candidate itemsets. Since each candidate is counted by only one process, all processes have to exchange database partitions during each iteration in order, for each process,

to obtain the global counts of the assigned candidate itemsets. The *Candidate Distribution* method also partitions candidate itemsets but replicates, instead of partition and exchanging, the database transactions. Thus, each process can proceed independently.

Vertical Mining: To better utilize the aggregate computing resources of parallel machines, a localized algorithm (Zaki *et. al*, 9) based on parallelization of *Eclat* was proposed and exhibited excellent scalability. It makes use of a vertical data layout by transforming the horizontal database transactions into vertical tid-lists of itemsets. By 1-item, the tid-list of an itemset is a sorted list of IDs for all transactions which contain the 1-itemset. Frequent k-itemsets are organized into disjoint equivalence classes by common $(k-1)$-prefixes, so that candidate $(k+1)$-itemsets can be generated by joining pairs of frequent k-itemsets from the same classes. The support of a candidate itemset can then be computed simply by intersecting the tid-lists of the two component subsets. Task parallelism is employed by dividing the mining tasks for different classes of itemsets among the available processes. The equivalence classes of all frequent 2-itemsets are assigned to processes and the associated tid-lists are distributed accordingly. Each process then mines frequent itemsets generated from its assigned equivalence classes independently, by scanning and intersecting the local tid-lists.

Pattern-Growth Method: The pattern-growth method derives frequent itemsets directly from the database without the costly generation and test of a large number of candidate itemsets. The detailed design is explained in the FP-growth algorithm (Han *et al.*, 10). Basically, it makes use of a frequent-pattern tree structure (FP-tree) where the repetitive transactions are compacted. Transaction itemsets are organized in that frequency-ordered prefix tree such that they share common prefix part as much as possible, and re-occurrences of items/itemsets are automatically counted. Then the FP-tree is pruned to mine all frequent patterns (itemsets). A partitioning-based, divide and conquer strategy is used to decompose the mining task into a set of smaller sub-tasks for mining confined patterns in the so-called conditional pattern bases. The conditional pattern base for each item is simply a small database of counted patterns that co-occur with the item. That small database is transformed into a conditional FP-tree that can be processed recursively (Zaiane *et al.*, 11, Pramudiono and Kitsuregawa 12, Li and Liu 13).

3.2 Decision Correlation Rules

Brin et al. (14) have proposed the extraction of correlation rules. The platform is no longer based on the support nor the confidence of the rules, but on the Chi-Squared statistical measure, written χ^2. The use of χ^2 is well-suited for several reasons: (i) it is a more significant measure in a statistical way than an association rule, (ii) the measure takes into account not only the presence but also the absence of the items and (iii) the measure is non-directional, and can thus highlight more complex existing links than a "*simple*" implication. Unlike

association rules, a correlation rule is not represented by an implication but by the patterns for which the value of the χ^2 function is larger than or equal to a given threshold.

Let r be a binary relation over a set of items $\mathcal{R} = \mathcal{I} \cup \mathcal{T}$. $\mathcal{I}$ represents the items of the binary relation used as analysis *criteria* and $\mathcal{T}$ is a target attribute. For a given transaction, the target attribute does not necessarily have a value. The computation of the value for the χ^2 function for an item $X \subseteq \mathcal{R}$ is based on its contingency table. In order to simplify the notation, we introduce, in a first step, the lattice of the literalsets associated with $X \subseteq \mathcal{R}$. This set contains all the literalsets that can be built up given X, and having $|X|$.

Definition 1 (Literalset Lattice). *Let $X \subseteq \mathcal{R}$ be a pattern, we denote by $\mathbb{P}(X)$ the literalset lattice associated with X. This set is defined as follows: $\mathbb{P}(X) = \{Y\overline{Z}$ such that $X = Y \cup Z$ and $Y \cap Z = \emptyset\} = \{Y\overline{Z}$ such that $Y \subseteq X$ and $Z = X \backslash Y\}$.*

Definition 2 (Contingency Table). *For a given pattern X, its contingency table, noted $CT(X)$, contains exactly $2^{|X|}$ cells. Each cell yields the support of a literalset $Y\overline{Z}$ belonging to the literalset lattice associated with X: the number of transactions including Y and containing no 1-item of Z.*

In order to compute the value of the χ^2 function for a pattern X, for each item $Y\overline{Z}$ belonging to its literalset lattice, we measure the difference between the square of the support of $Y\overline{Z}$ and its expectation value ($E(Y\overline{Z})$), and divide by the average of $Y\overline{Z}$ ($E(Y\overline{Z})$). Finally, all these values are summed.

$$\chi^2(X) = \sum_{Y\overline{Z} \in \mathbb{P}(X)} \frac{(Supp(Y\overline{Z}) - E(Y\overline{Z}))^2}{E(Y\overline{Z})}, \tag{1}$$

Brin et al. (14) have shown that there is a single degree of freedom between the items. A table giving the centile values in function of the χ^2 value for X can be used in order to obtain the correlation rate for X.

Definition 3 (Correlation Rule). *Let $MinCor$ (≥ 0) be a threshold given by the end-user and $X \subseteq \mathcal{R}$ a pattern. If $\chi^2(X) \geq MinCor$, then X is a valid correlation rule. If X contains an item of $\mathcal{T}$, then the obtained rule is called a Decision Correlation Rule (DCR).*

Moreover, in addition to the previous constraint, the Cochran *criteria* (Moore, 15) is used to evaluate whether a correlation rule is semantically valid: all literalsets of a contingency table must have an expectation value not equal to zero and 80% of them must have a support larger than 5% of the whole population. This last *criterium* has been generalized by Brin et al. (14) as follows: $MinPerc$ of the literalsets of a contingency table must have a support larger than $MinSup$, where $MinPerc$ and $MinSup$ are thresholds specified by the user.

Definition 4 (Equivalence Class associated with a literal). *Let $Y\overline{Z}$ be a literal. Let us denote by $[Y\overline{Z}]$ the equivalence class associated with the literal $Y\overline{Z}$. This class contains the set of transaction identifiers of the relation including Y and containing no value of Z (i.e., $[Y\overline{Z}] = \{i \in Tid(r)$ such that $Y \subseteq Tid(i)$ and $Z \cap Tid(i) = \emptyset\}$).*

Definition 5 (Contingency Vector). *Let $X \subseteq \mathcal{R}$ be a pattern. The contingency vector of X, denoted $CV(X)$, groups the set of the literalset equivalence classes belonging to $\mathbb{P}(X)$ ordered according to the lectic order.*

Since the union of the equivalence classes $[Y\overline{Z}]$ of the literalset lattice associated with X is a partition of the TIds, we ensure that a single transaction identifier belongs only to one single equivalence class. Consequently, for a given pattern X, its contingency vector is an exact representation of its contingency table. To derive the contingency table from a contingency vector, it is sufficient to compute the cardinality of each of its equivalence classes. If the literalsets, related to the equivalence classes of a CV, are ordered according to the lectic order, it is possible to know, because of the binary coding used, the literal relative to a position i of a contingency vector ($i \in [0; |X| - 1]$). This is because the literal and the integer i have the same binary coding. The following proposition shows how to compute the CV of the $X \cup A$ pattern given the CV of X and the set of identifiers of the relation containing pattern A.

Proposition 1. *Let $X \subseteq \mathcal{R}$ be a pattern and $A \in \mathcal{R} \backslash X$ a 1-item. The contingency vector of the $X \cup A$ pattern can be computed given the contingency vectors of X and A as follows:*

$$CV(X \cup A) = (CV(X) \cap [\overline{A}]) \cup (CV(X) \cap [A]) \tag{2}$$

In order to mine DCRs, we have proposed in (Casali and Ernst, 1) the LHS-Chi2 algorithm (see Alg. 1 for a simplified version). This algorithm is based both (*i*) on a double recursion in order to browse the search space according to the lectic order and (*ii*) on CVs.

The `CREATE_CV` function is used, given the contingency vector of a pattern X and the set of the transaction identifiers containing a 1-item A, to build the contingency vector of the pattern X ∪ A. The `CtPerc` predicate checks the relaxed Cochran *criteria*.

4 Extracting Correlated Patterns in Parallel

As it is easy to apprehend when regarding last section, the use of a recursive algorithm in a multicore programming environment is an effective challenge. This because recursion cannot be measured in terms of number of loops to perform. However, parallelizing existing algorithms is an important consideration as well. We first tried to replace the recursive calls by calls to appropriate threads, which quickly appeared as an impossible solution. Another possible approach was based

Alg. 1. LHS-Chi2 Algorithm

Input: X and Y two patterns
Output: $\{itemset\ Z \subseteq X$ such that $\chi^2(Z) \geq MinCor\}$
1: **if** $Y = \emptyset$ **and** $|X| \geq 2$ **and** $\exists t \in \mathcal{T} : t \in X$ **and** $\chi^2(X) \geq MinCor$ **then**
2: **Output** X, $\chi^2(X)$
3: **end if**
4: $A := max(Y)$
5: $Y := Y \backslash \{A\}$
6: LHS-Chi2(X,Y)
7: $Z := X \cup \{A\}$
8: VC(Z) := CREATE_CV(CV(X),Tid(A))
9: **if** $CtPerc(CV(Z), MinPerc, MinSup)$ **then**
10: LHS-Chi2(Z,Y)
11: **end if**

on the well known fact that each recursive algorithm can be rewritten in a iterative format. However, the *while* loop used to run over the used stack may not be evaluated in terms of a *for* loop due to the absence of explicit boundaries.

Finally, and in order to solution the problem, we recalled that we first compared our LHS-Chi2 algorithm to a Levelwise one, based on the same monotone and anti-monotone constraints but which did not include Contingency Vectors management. The main reason of the obtained performance gains is that pruning the search space using the lectic order is much more elegant than using the Levelwise order but has no impact nor on the results nor on the performances. On the other hand, generating the candidates at a given level is a bounded task, limited by the number of existing 1-items. This is why we decided to go back to the Levelwise order to prune in a parallel way the search space, and to keep the Contingency Vectors in order to manage the constraints.

The corresponding result is presented hereafter in the form of three functions. The overall algorithm, called *PLW_Chi2* (where PLW stands for Parallel Levelwise), demonstrates the parallel features of our method. The second function *dowork_level* constitutes each single Task executed in parallel. Finally, a particular aspect of such a Task is detailed in a third function, associated to specific shared resource management issues. We first present these functions through simplified code and comment them in a second stage.

```
void PLW_Chi2 (unsigned short X[], unsigned short sX, unsigned short sI)
// X[] : set of computed 1-items
// sX : number of valid 1-items within X[]
// sI : total number of 1-items in X[]
{
   unsigned long cit, nit; // number of candidates at level l and (l+1)
   cit = sX;
   for (unsigned char curlev = 2; curlev <= MaxLv && cit > 0; curlev++)
   {
       nit = 0L;
       T_Res aRes;
```

```
        combinable<unsigned long> lnit;
        parallel_for(0u, (unsigned) cit, [curlev, X, sX, &aRes, &lnit] (int i)
        {
            dowork_level (curlev, X, sX, i, sI, &aRes);
            lnit.local() += aRes.nit;  // ...
        });
        nit = lnit.combine(plus<unsigned long>());  // ...
        cit = nit;
        update_shared_resources ();
    }
}
```

Parallelization takes place at each level (*curlev* variable) of the LEVELWISE search algorithm. The number of launched Tasks at level l directly depends of the number of existing candidates at level *(l - 1)*, e.g. *cit*. Each Task corresponds principally to a call to the *dowork_level* function, which performs the work it is intended to do, and collects some statistics during the call through the *aRes* object. We interest us here only to a particular statistic, the *lnit* member of the *aRes* object, which sums the number of discovered candidates to be examined at the next level. Because each Task computes its own candidates for the next level, the method has to pay attention to the possible interference which could take place during the overall parallel computation on such a "shared" variable, which can be seen here as an aggregation pattern.

We use therefore a two-phase approach: First, we calculate partial results locally on a per-Task basis. Then, once all of the per-Task partial results are at disposal, we sequentially merge the results into one final accumulated value. The PPL provides a special data structure that makes it easy to create per-Task local results in parallel, and merge them as a final sequential step. This data structure is the *combinable* class. In the above code, the final accumulated object is the *lnit* object, which decomposes into local to each Task *lnit.local()* sub-objects. After the *parallel_for* loop achieves, the final sum is produced by invoking the *combine()* method on the global object.

The *dowork_level* function is for its part roughly implemented as follows:

```
void dowork_level (
    unsigned char nc, unsigned short pX[], unsigned short cX,
    unsigned long nel, unsigned short sIX, T_Res& pRes)
{
    unsigned short vmin, tCand[MaxLv + 1];  // a candidate
    unsigned long j, k;
    unsigned char *theVC;  // a CV
    // other declarations and initializations ...

    // get current itemset
    vmin = get_pattern (nc, tCand, pX, cX, nel, sIX);
    j = 0;   // get j, index of the first 1-item to add to the itemset
    while (j < cX && pX[j] <= vmin)  j++;
    for (k = j; k < cX; k++)
```

```
    {
      // add a 1-item to the current itemset to produce a candidate
      tCand[0] = pX[k];
      // compute its CV if the constraints are valid
      theVC = compute_CV (tCand, nc,  ...);
      // memorize the candidate and add it to results if applies
      store_CV (tCand, nc, theVC, pRes, ...);
      // update statistics
      (pRes.nit)++;  //...
    }
}
```

We shall not enter into the implementation details of this function. First because the code is most C likely and is easy to understand. And second because it does not include any specific parallel or shared memory features. So we shall only explain its overall functionalities. The *for* loop is used here to produce all the candidates of the current stage (the *tCand* variable). This is done by "adding" the possible existing 1-items to the base itemset managed by the function, and identified by the *nel* "number" (we shall discuss this aspect later). Once having generated such a candidate, we verify first if the different constraints underlying to our method are verified or not by the candidate. If it is the case, we compute its Contingency Vector (the whole is done by the *compute_CV* function). We second (try to) memorize the candidate in order to reuse it at the next level, and we add the candidate to the results if it contains one item of the target attribute.

The last function we present is *store_CV*. We focus moreover on a very specific section of code dedicated to the storage of results:

```
bool store_CV (unsigned short X[], unsigned short cardX, ...)
{
     // ... add X to the result file if X contains the target
     if (...)
     {
       critical_section cs;
       cs.lock();
       if (...)
         write_llhsp_to_file (X, cardX, ...);
       else
         write_pattern_to_file (X, cardX, ...);
       cs.unlock();
     }
   // ...
}
```

Let us first explain the functionality involved in the last *if* statement. k-itemsets verifying the whole defined constraints and including one item belonging to the target column have to be included into the results. This is managed through their insertion into data files (one is associated to each value of k).

During the parallellization process, each Task may write to one of these files each time it discovers a new valid itemset. What raises another shared resource problem, addressed by the PPL by the use of critical sections (a well-known concept in multi-threading developments), as shown in the above code. When encountering such an instruction at run-time, the OS will not authorize any other Task to execute before the "lock" has been released.

To finish this presentation, some explanations concerning the way we manage the memorization of candidates (and associated information such as Contingency Vectors). The main shared data structure in our developments is a tree storing the k-itemsets of "interest". The corresponding node structure, given in C:

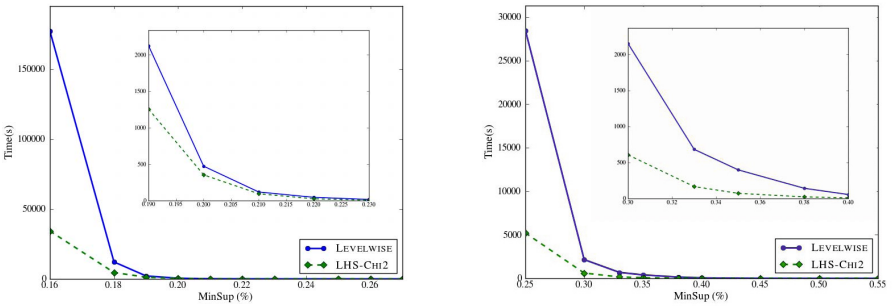
```
typedef struct pattern_node
{
   unsigned short *Mot; /* the pattern */
   unsigned char *pVC; /* pointer to the Contingency Vector */
   T_NM *frere; /* pointer to next node at same level */
   T_NM *fils; /* pointer to next node at lower level */
   ...
} T_NM;
```

Each time a Task discovers a candidate verifying the whole constraints, the candidate is inserted into the tree. The insertion by itself uses the critical section concept we just introduced. Because the stored itemsets (patterns) are lexicographically organized within the tree, each of them can be referred to by a node number (what explains the *nel* "number" introduced above). Finally, after evaluating the candidates, the exploring process will retain them or not. In the latter case, the tree structure may be garbaged, which is done by the *update_shared_resources* function called at the end of our global *PLW_Chi2*method.

5 Experimental Analysis

As briefly mentioned in the Introduction Section, this work has been initially applied on concrete data measurement files provided by two industrial manufacturing partners in the area of Microelectronics : STMicroelectronics (STM) and ATMEL (ATM, which became LFoundry). The results of the realized experimental series are presented on 2 plans to be followed. They are associated with an analysis of 2 files among those supplied by both manufacturers. The first one (STM) contains 1241 columns and 296 lines. The second (ATM) consists of 749 columns and 213 lines. We chose a target attribute among a few possible columns. In both cases, the presented diagrams show the execution times of two methods when $MinSup$ varies while $MinPerc$ (0.34 for the STM file and 0.24 for the ATM one) and $MinCor$ (1.6 resp. 2.8) are fixed. The signification of these parameters were given in section 3.2.

Figures 1(a) and 1(b), extracted from (Casali and Ernst, 1), show the execution times of a standard Levelwise algorithm and the LHS-Chi2 algorithm on a non core computer (a HP Workstation with a 1.8 GHz processor and 4 Gb



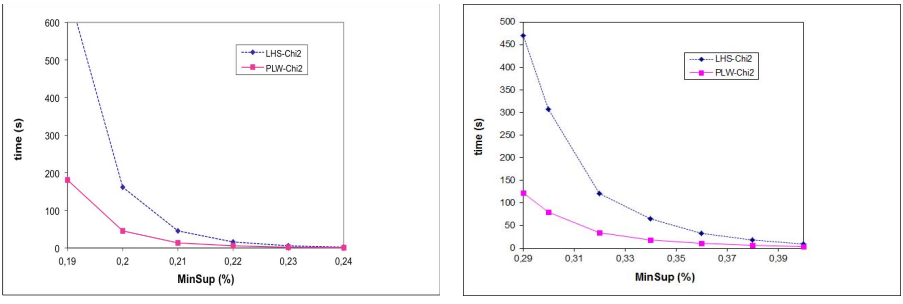
(a) Results for STM File (b) Results for ATM File

Fig. 1. Execution times with a single processor

RAM, working under a Windows XP 32 bits OS). The difference between the two methods is that the LEVELWISE method uses no contingency vectors but standard computation of contingency tables. As the graphs point it out, the response times of the LHS-CHI2 method are between 30% and 70% better than LEVELWISE. An increasing windowing of the results is provided for subsequent sub-intervals of *MinSup*.

Figures 2(a) and 2(b) show the same execution times using the LHS-CHI2 algorithm and the presented PLW-CHI2 algorithm on a 4 core computer (a DELL Workstation with a 2.8 GHz processor and 12 Gb RAM working under the Windows 7 64 bits OS).

As it is easy to understand, the LHS-CHI2 method works here about two times faster on the multicore architecture, this not because of the number of cores (which are not used) but because of the computer basic enhanced capabilities. When regarding to the performances of the PLW-CHI2 method, there is a gain factor of about 3.5, which is to compare to the number of available



(a) Results for STM File (b) Results for ATM File

Fig. 2. Execution times with 4 cores

cores, which is 4. In other words, the parallelization of the LHS-Chi2 algorithm raises performance gains practically equals to the number of cores, the (little) loss being due to the shared memory management issues.

6 Conclusion and Future Work

In this paper, we present a new approach to discover correlated patterns, based on the usage of multicore architectures. Our approach is based on two concepts: Contingency Vectors, an alternate representation of contingency tables, and the Parallel Patterns Library (PPL). One of the advantages of Contingency Vectors is that they allow the Chi-squared computation of a k-itemset directly from one of its subsets. However, the usage of the PPL has two disadvantages: on one hand we need to use lambda calculus, and on the other hand, the parallelization of recursive algorithms is hard (we do not control neither the number of cores, nor the depth of the tree), even if we derecursify the algorithm. That is why we have chosen to implement a Levelwise algorithm which implements these two concepts. Experiments are convincing because our new algorithm obtains a time gain factor of about 3.5 (when using 4 cores) in comparison with the recursive version.

For future works, we intend to develop a new version of the recursive algorithm using a bitmap representation for a Contingency Vector, thus we can minimize disk I/O, and, for finer control processors, build our own thread manager.

References

Casali, A., Ernst, C.: Discovering correlated parameters in semiconductor manufacturing processes: A data mining approach. IEEE Transactions on Semiconductor Manufacturing 25(1), 118–127 (2012)

Darlington, J., Ghanem, M.: ke Guo, Y., To, H.W.: Guided resource organisation in heterogeneous parallel computing (1996)

Tatikonda, S., Parthasarathy, S.: Mining tree-structured data on multicore systems. PVLDB 2(1), 694–705 (2009)

Mitchell, J.C.: Foundations for programming languages. Foundation of computing series. MIT Press (1996)

Herlihy, M., Shavit, N.: The art of multiprocessor programming. Morgan Kaufmann (2008)

Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Bocca, J.B., Jarke, M., Zaniolo, C. (eds.) VLDB, pp. 487–499. Morgan Kaufmann (1994)

Agrawal, R., Shafer, J.C.: Parallel mining of association rules. IEEE Trans. Knowl. Data Eng. 8(6), 962–969 (1996)

Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2000)

Zaki, M.J., Parthasarathy, S., Ogihara, M., Li, W.: Parallel algorithms for discovery of association rules. Data Min. Knowl. Discov. 1(4), 343–373 (1997)

Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Chen, W., Naughton, J.F., Bernstein, P.A. (eds.) SIGMOD Conference, pp. 1–12. ACM (2000)

Zaïane, O.R., El-Hajj, M., Lu, P.: Fast parallel association rule mining without candidacy generation. In: Cercone, N., Lin, T.Y., Wu, X. (eds.) ICDM, pp. 665–668. IEEE Computer Society (2001)

Pramudiono, I., Kitsuregawa, M.: Tree structure based parallel frequent pattern mining on PC cluster. In: Mařík, V., Štěpánková, O., Retschitzegger, W. (eds.) DEXA 2003. LNCS, vol. 2736, pp. 537–547. Springer, Heidelberg (2003)

Li, E., Liu, L.: Optimization of frequent itemset mining on multiple-core processor. In: VLDB, pp. 1275–1285 (2007)

Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: Generalizing association rules to correlations. In: SIGMOD Conference, pp. 265–276 (1997)

Moore, D.: Measures of lack of fit from tests of chi-squared type. Journal of Statistical Planning and Inference 10(2)(2), 151–166 (1984)

Index Data Structure for Fast Subset and Superset Queries

Iztok Savnik

[1] Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, 6000 Koper
[2] Artificial Intelligence Laboratory, Jozef Stefan Institute, 1000 Ljubljana, Slovenia
iztok.savnik@upr.si

Abstract. A new data structure *set-trie* for storing and retrieving sets is proposed. Efficient manipulation of sets is vital in a number of systems including datamining tools, object-relational database systems, and rule-based expert systems. Data structure *set-trie* provides efficient algorithms for set containment operations. It allows fast access to subsets and supersets of a given parameter set. The performance of operations is analyzed empirically in a series of experiments on real-world and artificial datasets. The analysis shows that sets can be accessed in $\mathcal{O}(c * |set|)$ time where $|set|$ represents the size of parameter set and c is a constant.

Keywords: subset queries, set containment queries, partial matching, access methods, database index.

1 Introduction

Set containment queries are common in various systems including datamining tools, object-relational databases, rule-based expert systems, and AI planning systems. Enumeration of subsets of a given universal set U is very common in *data mining* algorithms [10] where sets are used as basis for the representation of hypotheses and search space forms a lattice. Often we have to see if a given hypothesis has already been considered by the algorithm. This can be checked by searching the set of hypotheses (sets) that have already been processed. Furthermore, in some cases hypotheses can be easily overthrown if a superset hypothesis has already been shown not valid. Such problems include discovery of association rules, functional dependencies as well as some forms of propositional logic [10,15,5,8].

In *object-relational database management systems* tables can have set-valued attributes i.e. attributes that range over sets. Set containment queries can express either selection or join operation based on set containment condition. Efficient access to relation records based on conditions that involve set operations are vital for fast implementation of such queries [12,18,7].

Rule-based expert systems use set containment queries to implement fast pattern-matching algorithms that determine which rules are fired in each cycle of expert system execution. Here sets form pre-conditions of rules composed of elementary conditions.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 134–148, 2013.

Given a set of valid conditions the set of fired rules includes those with pre-condition included in this set [6,4].

Finally, in *AI planning systems* goal sets are used to store goals to be achieved from a given initial state. Planning modules use subset queries in procedure that examines if a given goal set is satisfiable. Part of the procedure represents querying goal sets that were previously shown to be unsatisfiable. Here also sets are used to form basic structure of hypothesis space [2].

In this paper we propose a novel index data structure *set-trie* that implements efficiently basic two types of set containment queries: *subset* and *superset queries*. *Set-trie* provides storage for sets as well as multisets. Preliminary version of this paper has been published in [16].

Set-trie is a tree data structure similar to $trie$ [13]. The possibility to extend the performance of usual $trie$ from membership operation to subset and superset operations comes from the fact that we are storing *sets* (*multisets*) and not the *sequences* of symbols as for ordinary tries. In case of sets (multisets) ordering of symbols in a set is not important as it is in the case of text. As it will be presented in the paper, the ordering of set elements is used as the basis for the definition of efficient algorithms for set containment operations. Since the semantics of set containment operations is equivalent to the semantics of multiset containment operations we will in the following text sometimes refer to both, sets and multisets, as *sets*.

We analyze subset and superset operations in two types of experiments. Firstly, we examine the execution of the operations on real-world data where multisets represent words from the English dictionary. Secondly, we have tested the operations on artificially generated data. In these experiments we tried to see how three main parameters: the size of sets, the size of *set-trie* tree and the size of test-set, affect the behavior of the operations. Analysis shows that sets can be accessed in $\mathcal{O}(c * |set|)$ time where $|set|$ represents the size of parameter set and c is a constant. The constant c is up to 5 for subset case and approximately 150 in average case for the superset case.

The paper is organized as follows. The following section presents the data structure *set-trie* together with the operations for searching the subsets and supersets in a tree. The Section 3 describes the empirical study of *set-trie*. We present a series of experiments that measure the behavior of operations and the size of data structure. Related work is presented in Section 4. We give presentation of existent work from the fields of algorithms and data structures, AI systems where sets are used for querying hypotheses and states, and object-relational database systems where indexes are used to access set-valued attributes. Finally, the conclusions and the directions of our further work are given in Section 5.

2 Data Structure *set-trie*

Set-trie is a tree composed of nodes labeled with indices from 1 to N where N is the size of the alphabet. The root node is labeled with $\{\}$ and its children can be the nodes labeled from 1 to N. A root node alone represents an empty set. A node labeled i can have children labeled with numbers greater than or equal i. Each node can have a flag denoting the last element in the set. Therefore, a set is represented by a path from the root node to a node with flag set to true.

Let us give an example of *set-trie*. Figure 2 presents a *set-trie* containing the sets $\{1,3\}, \{1,3,5\}, \{1,4\}, \{1,2,4\}, \{2,4\}, \{2,3,5\}$. Note that flagged nodes are represented with circles.

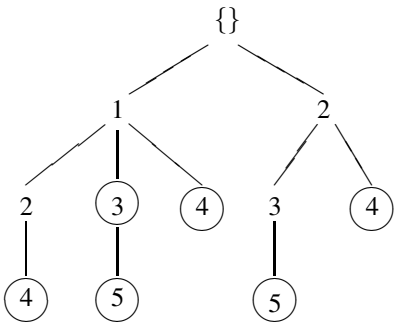


Fig. 1. Example of *set-trie*

Since we are dealing with sets for which the ordering of the elements is not important, we can define a syntactical order of symbols by assigning each symbol a unique index. Words are sequences of symbols ordered by indices. The ordering of symbols is exploited for the *representation* of sets of words as well as in the *implementation* of the above stated operations.

Set-trie is a tree storing a set of words which are represented by a path from the root of *set-trie* to a node corresponding to the indices of elements from words. As with tries, prefixes that overlap are represented by a common path from the root to an internal vertex of *set-trie* tree.

The operations for searching subsets and supersets of a set X in S use the ordering of U. The algorithms do not need to consider the tree branches for which we know they do not lead to results. The search space for a given X and tree representing S can be seen as a subtree determined primarily by the search word X but also with the search tree corresponding to S.

2.1 Set Containment Operations

Let us first give formal definition of the basic subset and superset operations. Let U be a set of ordered symbols. The subsets of U are denoted as *words*. Given a set of words S and a subset of U named X, we are interested in the following queries.

1) existsSubset(S, X) returns $true$ if $\exists Y \in S : Y \subseteq X$ and $false$ otherwise.
2) existsSuperset(S, X): returns $true$ if $\exists Y \in S : X \subseteq Y$ and $false$ otherwise.
3) getAllSubsets(S, X): returns all sets Y such that $Y \in S \wedge Y \subseteq X$.
4) getAllSupersets(S, X): returns all sets Y such that $Y \in S \wedge X \subseteq Y$.

Let us now present a data structure *Word* for storing sets of symbols. Symbols in words are represented as integer numbers. Elements of a set represented by $Word$ can be scanned using the following mechanism. The operation $word.gotoFirstElement$

sets the current element of word to the first element of ordered set. Then, the operation $word.existsCurrentElement$ checks if word has the current element set. The operation $word.currentElement$ returns the current element, and the operation $word.gotoNextElement$ goes to the next element in the set.

Using data structure $Word$ we can now describe the operations of the data structure *set-trie*. The first operation is *insertion*. The operation $insert$(root,word) enters a new $word$ into the *set-trie* referenced by the root $node$. The operation is presented by Algorithm 1.

Algorithm 1. insert($node$, $word$)

```
 1: if (word.existsCurrentElement) then
 2:     if (exists child of node labeled word.currentElement) then
 3:         nextNode = child of node labeled word.currentElement;
 4:     else
 5:         nextNode = create child of node labeled word.currentElement;
 6:     end if
 7:     insert(nextNode, word.gotoNextElement)
 8: else
 9:     node's flag_last = true;
10: end if
```

Each invocation of operation $insert$ either traverses through the existing tree nodes or creates new nodes to construct a path from the root to the flagged node corresponding to the last element of the ordered set.

The following operation $search$(node,word) searches for a given $word$ in the tree $node$. It returns true when it finds all symbols from the word, and false as soon one symbol is not found. The algorithm is shown in Algorithm 2. It traverses the tree $node$ by using the elements of ordered set $word$ to select the children.

Algorithm 2. search($node$, $word$)

```
 1: if (word.existsCurrentElement) then
 2:     if (there exists child of node labeled word.currentElement) then
 3:         matchNode = child vertex of node labeled word.currentElement;
 4:         search(matchNode, word.gotoNextElement);
 5:     else
 6:         return false;
 7:     end if
 8: else
 9:     return (node's last_flag == true) ;
10: end if
```

Let us give a few comments to present the algorithm in more detail. The operation have to be invoked with the call $search$(root,word.gotoFirstElement) so that $root$ is

the root of the *set-trie* tree and the current element of the $word$ is the first element of $word$. Each activation of $search$ tries to match the current element of $word$ with the child of $node$. If the match is not successful it returns $false$ otherwise it proceeds with the following element of $word$.

The operation $existsSubset$(node,word) checks if there exists a subset of $word$ in the given tree referenced by $node$. The subset that we search in the tree has fewer elements than $word$. Therefore, besides that we search for the exact match we can also skip one or more elements in $word$ and find a subset that matches the rest of the elements of $word$. The operation is presented in Algorithm 3.

Algorithm 3. existsSubset(node,set)

```
 1: if (node.last_flag == true) then
 2:     return true;
 3: end if
 4: if (not word.existsCurrentElement) then
 5:     return false;
 6: end if
 7: found = false;
 8: if (node has child labeled word.currentElement) then
 9:     nextNode = child of node labeled word.currentElement;
10:     found = existsSubset(nextNode, word.gotoNextElement);
11: end if
12: if (!found) then
13:     return existsSubset(node,word.gotoNextElement);
14: else
15:     return true;
16: end if
```

Algorithm 3 tries to match elements of $word$ by descending simultaneously in tree and in $word$. The first IF statement (line 1) checks if a subset of $word$ is found in the tree i.e. the current node of a tree is the last element of subset. The second IF statement (line 4) checks if $word$ has run of the elements. The third IF statement (line 8) verifies if the parallel descend in $word$ and tree is possible. In the positive case, the algorithm calls $existsSubset$ with the next element of $word$ and a child of $node$ corresponding to matched symbol (parallel descend). Finally, if match did not succeed, current element of $word$ is skipped and $existsSubset$ is called with same $node$ and next element of $word$ in line 13.

The operation $existsSubset$ can be easily extended to find all subsets of a given $word$ in a tree $node$. After finding the subset in line 15 the subset is stored and the search continues in the same manner as before. The experimental results with the operation $getAllSubsets$(node,word) are presented in the following section.

The operation $existsSuperset$(node,word) checks if there exists a superset of $word$ in the tree referenced by $node$. While in operation $existsSubset$ we could skip some elements from $word$, here we can do the opposite: the algorithm can skip some elements in supersets represented by $node$. Therefore, $word$ can be matched with the subset of superset from a $tree$. The operation is presented in Algorithm 4.

Algorithm 4. existsSuperset($node$, $word$)

```
 1: if (not word.existsCurrentElement) then
 2:     return true;
 3: end if
 4: found = false;
 5: from = word.currentElement;
 6: upto = word.nextElement if it exists and N otherwise;
 7: for (each child of node labeled l: from < l ≤ upto) & (while not found) do
 8:     if (child is labeled upto) then
 9:         found = existsSuperset(child,word.gotoNextElement);
10:     else
11:         found = existsSuperset(child,word);
12:     end if
13: end for
```

Let us present Algorithm 4 in more detail. The first IF statement checks if we are already at the end of $word$. If so, then the parameter $word$ is covered completely with a superset from $tree$. Lines 5-6 set the lower and upper bounds of iteration. In each pass we either take current $child$ and call $existsSuperset$ on unchanged $word$ (line 11), or, descend in parallel on both $word$ and tree in the case that we reach the upper bound i.e. the next element in $word$ (line 9).

Again, the operation $existsSuperset$ can be quite easily extended to retrieve all supersets of a given $word$ in a tree $node$. However, after $word$ (parameter) is matched completely (line 2 in Algorithm 4), there remains a subtree of trailers corresponding to a set of supersets that subsume $word$. This subtree is rooted in a tree node, let say $node_k$, that corresponds to the last element of $word$. Therefore, after the $node_k$ is matched against the last element of the set in line 2, the complete subtree has to be traversed to find all supersets that go through $node_k$.

3 Experiments

The performance of the presented operations is analyzed in four experiments. The main parameters of experiments are: number of words in tree, size of the alphabet, and maximal length of words. The parameters are named: $numTreeWord$, $alphabetSize$, and $maxSizeWord$, respectively. In every experiment we measure the *number of visited nodes necessary for an operation to terminate*.

In the first experiment, *set-trie* is used to store real-world data – it stores multisets obtained from words of English Dictionary. In the following three experiments we use artificial data – datasets include randomly generated sets of sets. In these experiments we analyze in detail the interrelations between one of the stated tree parameters and the number of visited nodes.

In all experiments we observe four operations presented in the previous section: $existsSubset$ (abbr. esb) and its extension $getAllSubsets$ (abbr. gsb), and $existsSuperset$ (abbr. esr) and its extension $getAllSupersets$ (abbr. gsr).

3.1 Experiment with Real-World Data

Let us now present the first experiment in more detail. The number of words in test set is 224,712 which results in a tree with 570,462 nodes. The length of words are between 5 and 24 and the size of the alphabet ($alphabetSize$) is 25. The test set contains 10,000 words.

Results are presented in Table 1 and Figure 2. Since there are 10,000 words and 23 different word lengths in the test set, approximately 435 input words are of the same length. Table 1 and Figure 2 present the average number of visited nodes for each input word length (except for gsr where values below word length 6 are intentionally cut off).

word length	esr	gsr	esb	gsb
2	523	169694	1	1
3	3355	103844	3	3
4	12444	64802	6	6
5	9390	34595	11	12
6	11500	22322	14	19
7	12148	17003	18	32
8	8791	10405	19	46
9	6985	7559	19	78
10	3817	3938	21	102
11	3179	3201	20	159
12	2808	2820	20	221
13	2246	2246	22	290
14	1651	1654	19	403
15	1488	1488	18	575
16	895	895	19	778
17	908	908	20	925
18	785	785	18	1137
19	489	489	22	1519
20	522	522	19	1758
21	474	474	19	2393
22	399	399	17	3044
23	362	362	17	3592
24	327	327	19	4167

Fig. 2. Visited nodes for dictionary words

Let us give some comments on the results presented in Table 2. First of all, we can see that the superset operations (esr and gsr) visit more nodes than subset operations (esb and gsb).

The number of nodes visited by esr and gsr decreases as the length of words increases. This can be explained by more constrained search in the case of longer words, while it is very easy to find supersets of shorter words and, furthermore, there are a lot of supersets of shorter words in the tree.

Since operation gsr returns all supersets (of a given set), it always visits more nodes than the operation esr. However, searching for the supersets of longer words almost

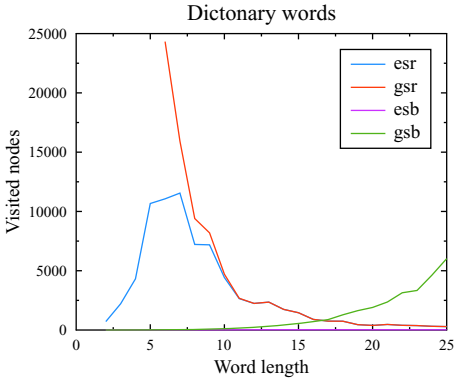


Fig. 3. Number of visited nodes

always results in failure and for this reason the number of visited nodes is the same for both operations.

The number of visited nodes for esb in the case that words have more than 5 symbols is very similar to the length of words. Below this length of words both esb and gsb visit the same number of nodes, because there were no subset words of this length in the tree and both operations visit the same nodes.

The number of visited nodes for gsb linearly increases as the word length increases. We have to visit all the nodes that are actually used for the representation of all subsets of a given parameter set.

3.2 Experiments with Artificial Data

Three experiments were done by using artificially generated data. Experiments are named $experiment1$, $experiment2$ and $experiment3$.

1) In $experiment1$ we observe relation between maximal length of words to number of visited nodes of *set-trie* in all four operations.
2) $experiment2$ shows the relation between number of words stored in *set-trie* and number of visited nodes in *set-trie* in all four operations.
3) $experiment3$ investigates the relation between the size of alphabet and number of visited nodes of *set-trie* in all four operations.

Let us start with $experiment1$. Here we observe the influence of maximal length of words to the performance of all four operations. We created four trees with $alphabetSize$ 30 and $numTreeWord$ 50,000. $maxSizeWord$ is different in each tree: 20, 40, 60 and 80, for tree1, tree2, tree3 and tree4, respectively. The length of word in each tree is evenly distributed between the minimal and maximal word size. The number of nodes in the trees are: 332,182, 753,074, 1,180,922 and 1,604,698. The test set contains 10,000 words.

Figure 3 shows the performance of all four operations on all four trees. The performance of superset operations is affected more by the change of word length than the subset operations.

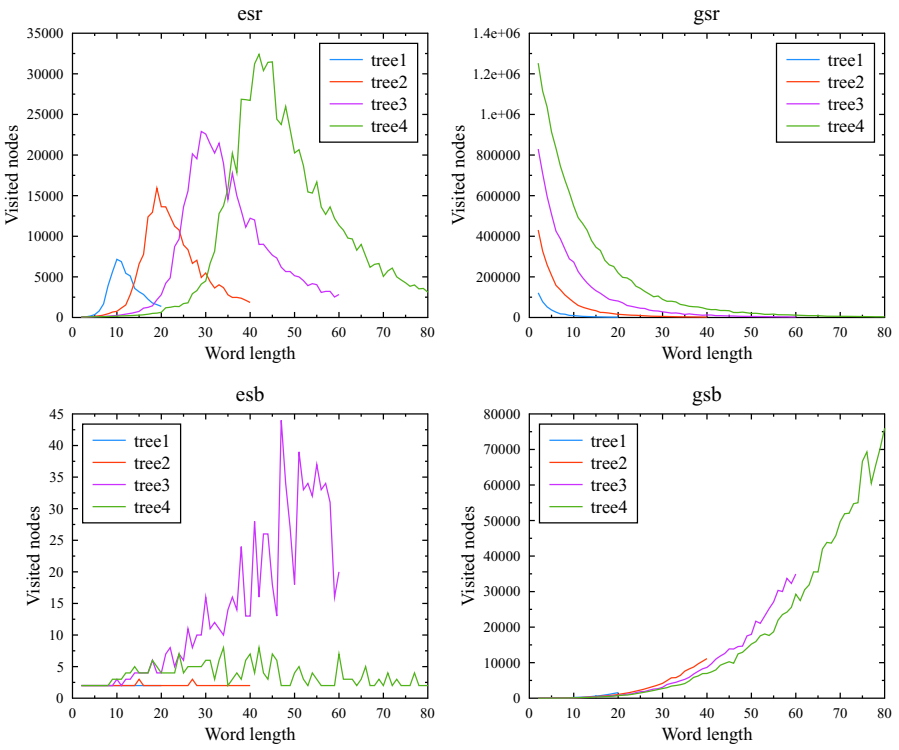


Fig. 4. Experiment 1 - increasing $maxSizeWord$

With an even distribution of data in all four trees, esr visits most nodes for input word lengths that are about half of the size of $maxSizeWord$ (as opposed to dictionary data where it visits most nodes for word lengths approximately one fifth of $maxSizeWord$). For word lengths equal to $maxSizeWord$ the number of visited nodes is roughly the same for all trees, but that number increases slightly as the word length increases.

esb operation visits fewer than 10 nodes most of the time, but for $tree3$ it goes up to 44. The experiment was repeated multiple (about 10) times, and in every run the operation jumped up in a different tree. As will be seen later in $experiment2$, it seems that $numTreeWord$ 50 is just on the edge of the value where esb stays constantly below 10 visited nodes. It is safe to say that the change in $maxSizeWord$ has no major effect on $existsSubSet$ operation.

In contrast to gsr, gsb visits less nodes for the same input word length in trees with greater $maxSizeWord$, but the change is minimal. For example for word length 35 in $tree2$ ($maxSizeWord$ 40) gsb visits 7,606 nodes, in $tree3$ ($maxSizeWord$ 60) it visits 5,300 nodes and in $tree4$ ($maxSizeWord$ 80) it visits 4,126 nodes.

In $experiment2$ we are interested about how a change in the number of words in the tree affects the operations. Ten trees are created all with $alphabetSize$ 30 and $maxSizeWord$ 30. $numTreeWord$ increases in each subsequent tree by 10,000 words: $tree1$ has 10,000 words, and $tree10$ has 100,000 words. The number of nodes

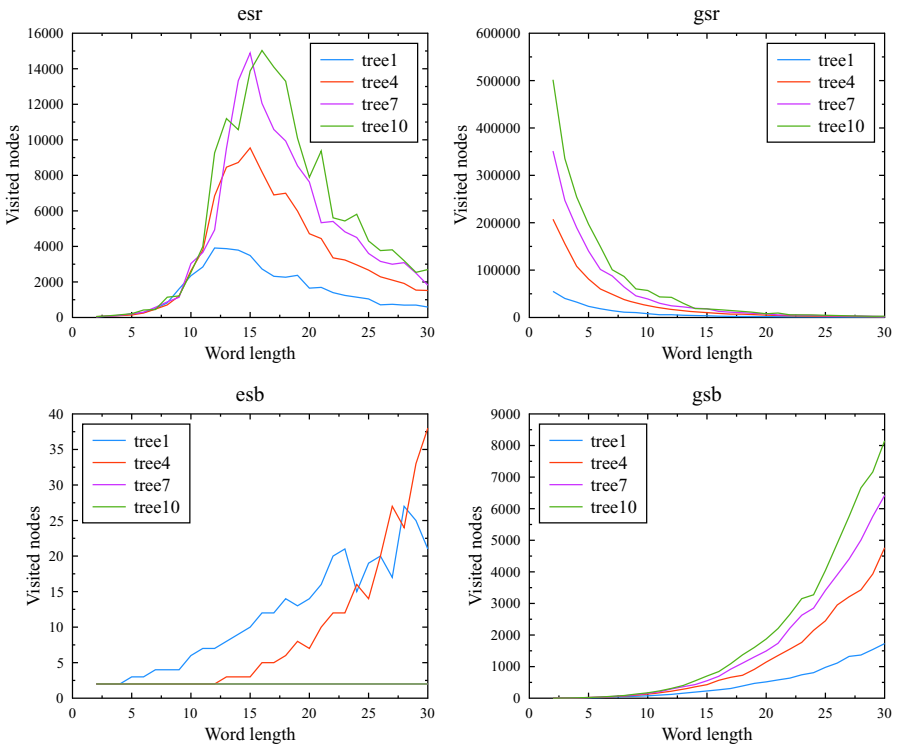


Fig. 5. Experiment 2 - increasing $numTreeWord$

in the trees (from $tree1$ to $tree10$) are: 115,780, 225,820, 331,626, 437,966, 541,601, 644,585, 746,801, 846,388, 946,493 and 1,047,192. The test set contains 5,000 words.

Figure 4 shows the number of visited nodes for each operation on four trees: $tree1$, $tree4$, $tree7$ and $tree10$ (only every third tree is shown to reduce clutter). When increasing $numTreeWord$ the number of visited nodes increases for esr, gsr and gsb operations. esb is least affected by the increased number of words in the tree. In contrast to the other three operations, the number of visited nodes decreases when $numTreeWord$ increases.

For input word lengths around half the value of $maxSizeWord$ (between 13 and 17) the number of visited nodes for esr increases with the increase of the number of words in the tree. For input word lengths up to 10, the difference between trees is minimal. After word lengths about 20 the difference in the number of visited nodes between trees starts to decline. Also, trees 7 to 10 have very similar results. It seems that after a certain number of words in the tree the operation remains constant.

The increased number of words in the tree affects the gsr operation mostly in the first quarter of $maxSizeWord$. The longer the input word, the lesser the difference between trees. Still, this operation is affected most by the change of $numTreeWord$. The average number of visited nodes for all input word lengths in tree1 is 8,907 and in

tree10 it is 68,661. Due to the nature of operation, this behavior is expected. The more words there are in the tree, the more supersets can be found for an input word.

As already noted above, when the number of words in the tree increases, number of visited nodes for esb decreases. After a certain number of words, in our case this was around 50,000, the operation terminates with minimal (possible) number of visited nodes for any word length. The increase of $numTreeWord$ pushes down the performance of operation (from left to right). This can be seen in Figure 4 by comparing $tree1$ and $tree4$. In $tree1$ the operation visits more then 10 after word length 15, and in $tree4$ it visits more than 10 nodes after word length 23. Overall the number of visited nodes is always low.

The chart of gsb operation looks like a mirrored chart of gsr. The increased number of words in tree has more effect on input word lengths where the operation visits more nodes (longer words). Below word length 15 the difference between trees is in the range of 100 visited nodes. At word length 30 gsb visits 1,729 nodes in $tree1$ and 8,150 nodes in $tree10$. Explanation of the increased number of visited nodes is similar as for gsr operation: the longer the word, the more subsets it can have, the more words in the tree, the more words with possible subsets there are.

In $experiment3$ we are interested about how a change of alphabet size affects the operations. Five trees are created with $maxSizeWord$ 50 and $numTreeWord$ 50,000. $alphabetSize$ is 20, 40, 60, 80 and 100, for $tree1$, $tree2$, $tree3$, $tree4$ and $tree5$, respectively. The number of nodes in the trees are: 869,373, 1,011,369, 1,069,615, 1,102,827 and 1,118,492. The test set contains 5,000 words.

When increasing $alphabetSize$ the tree becomes sparser–the number of child nodes of a node is larger, but the number of nodes in all five trees is roughly the same. Operation gsr and more notably gsb operation, visit less nodes for the same input word length: the average number of visited nodes decreased when $alphabetSize$ increases. Operation esr on the other hand visits more nodes in trees with larger alphabetSize.

Number of visited nodes of esr increases with the increase of $alphabetSize$. This is because it is harder to find supersets of given words, when the number of symbols that make up words is larger. This is more evident for word lengths below half $maxSizeWord$. The number of visited nodes starts decreasing rapidly after a certain word length. At this point the operation does not find any supersets and it returns false.

Operation gsr is not affected much by the change of $alphabetSize$. More evident change appears when $alphabetSize$ is increased over 20 ($tree1$). The number of visited nodes in trees 2 to 5 is almost the same, but it does decrease with the increase of $alphabetSize$.

In $tree1$ esb visits on average 3 nodes. When we increase $alphabetSize$ the number of visited nodes also increases, but as in gsr the difference between trees 2 to 5 is small.

The change of $alphabetSize$ has more significant effect on longer input words for the gsr operation. The number of visited nodes decreased when $alphabetSize$ increased. Here again the most evident change is when going over $alphabetSize$ 20. In each subsequent increase, the difference in the number of visited nodes is smaller.

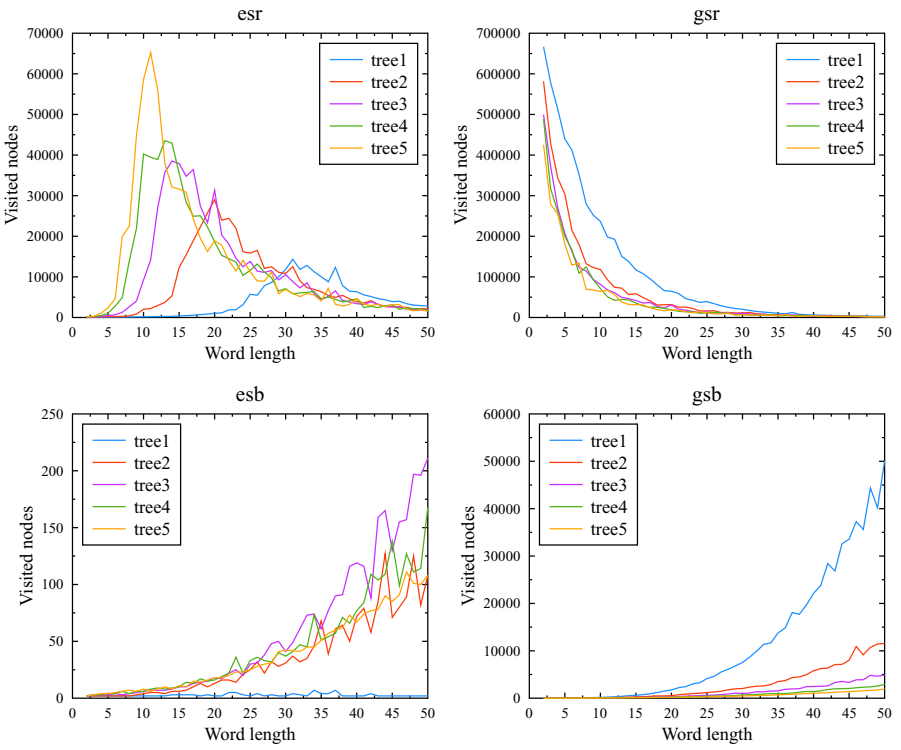


Fig. 6. Experiment 3 - increasing $alphabetSize$

4 Related Work

The problem of querying sets of sets appears in various areas of Computer Science. Firstly, the problem has been studied in the form of *substring search* by Rivest [13], Baeza-Yates [1] and Charikar [3]. Secondly, the subset queries are studied in various sub-areas of AI for storing and querying: pre-conditions of a large set of rules [6], states in planning for storing goal sets [8] and hypotheses in data mining algorithms [9]. Finally, querying sets is an important problem in object-relational databases management systems where attributes of relations can range over sets [18,12,7,19,20].

4.1 Partial-Matching and Containment Query Problem

The data structure we propose is similar to trie [13,14]. Since we are not storing sequences but *sets* we can exploit the fact that the order in sets is not important. Therefore, we can take advantage of this to use syntactical order of elements of sets and obtain additional functionality of tries.

Our problem is similar to searching substrings in strings for which *tries* and *Suffix trees* can be used. Firstly, Rivest examines [13] the problem of partial matching with the

use of hash functions and trie trees. He presents an algorithm for partial match queries using tries. However, he does not exploit the ordering of indices that can only be done in the case that *sets* or *multisets* are stored in tries.

Baeza-Yates and Gonnet present an algorithm [1] for searching regular expressions using Patricia trees as the logical model for the index. They simulate a finite automata over a binary Particia tree of words. The result of a regular expression query is a superset or subset of the search parameter.

Finally, Charikar et. al. [3] present two algorithms to deal with a subset query problem. The purpose of their algorithms is similar to $existsSuperSet$ operation. They extend their results to a more general problem of orthogonal range searching, and other problems. They propose a solution for "containment query problem" which is similar to our 2. query problem introduced in Section 1.

4.2 Querying Hypotheses and States in AI Systems

The initial implementation of *set-trie* was in the context of a datamining tool $fdep$ which is used for the induction of functional from relations [15,5]. It has been further used in datamining tool $mdep$ for the induction of multivalued dependencies from relations [17]. In both cases sets are used as the basis for the representation of dependencies. Hypotheses (dependencies) are checked against the negative cover of *invalid dependencies* represented by means of *set-trie*. Furthermore, positive cover including valid dependencies is minimized by using *set-trie* as well.

Doorenbos in [4] proposes an index structure for querying pre-conditions of rules to be matched while selecting the next rule to activate in a rule-based system Rete [6]. Index structure stores conditions in separate nodes that are linked together to form pre-conditions of rules. Common conditions of rules are shared among the rules: lists of conditions with common prefix share all nodes that form prefix. Given a set of conditions that are fulfilled all rules that contain as pre-condition a subset of given set of conditions can be activated.

Similar index structure is proposed by Hoffman and Koehler by means of Unlimited Branching Tree (abbr. UBTree) to store set of sets. The main difference with the representation of rules in expert systems is that UBTree does not use variables. Children of node are stored in a list attached to node. A set is in UBTree represented by a path from root to final node; path is labeled by elements of a set. The search procedures for subset and superset problems are similar to those we propose, however, the main difference in procedures is that we explicitly use ordering of sets for search while Hoffman and Koehler give a more general algorithm allowing other heuristic to be exploited. Our publication in 1993 [15] evidently presents the independence of work.

4.3 Indexing Set-Valued Attributes of Object-Relational Databases

Sets are among important data modeling constructs in object-relational and object-oriented database systems. *Set-valued attributes* are used for the representation of properties that range over sets of atomic values or objects. Database community has shown significant interest in indexing structures that can be used as access paths for querying set-valued attributes [18,12,7,19,20].

Set containment queries were studied in the frame of different index structures. Helmer and Moercotte investigated four index structures for querying set-valued attributes of low cardinality [7]. All four index structures are based on conventional techniques: signatures and inverted files. Index structures compared are: sequential signature files, signature trees, extendable signature hashing, and B-tree based implementation of inverted lists. Inverted file index showed best performance over other data structures in most operations.

Zhang et al. [20] investigated two alternatives for the implementation of containment queries: a) separate IR engine based on inverted lists and b) native tables of RDBMS. They have shown that while RDBMS are poorly suited for containment queries they can outperform inverted list engine in some conditions. Furthermore, they have shown that with some modifications RDBMS can support containment queries much more efficiently.

Another approach to the efficient implementation of set containment queries is the use of signature-based structures. Tousidou et al. [19] combine the advantages of two access paths: linear hashing and tree-structured methods. They show through the empirical analysis that S-tree with linear hash partitioning is efficient data structure for subset and superset queries.

5 Conclusions

The paper presents a data structure *set-trie* that can be used for efficient storage and retrieval of subsets or supersets of a given $word$. The algorithms of set containment operations are analyzed empirically. It has been demonstrated that the algorithms are stable when used on real-world and artificially generated data. Empirical analysis was used to determine the behavior of each particular set containment operations. The performance of *set-trie* is shown to be efficient enough for storage and retrieval of sets in practical applications.

Initial experiments have been done to investigate if *set-trie* can be employed for searching substrings and superstrings in texts. For this purpose the data structure *set-trie* has to be augmented with the references to the position of words in text. As in the case of indexes used in information retrieval [11] *set-trie* can be decomposed into *dictionary* and *postings*. Empirical analysis which would show memory consumption and efficiency of *set-trie* used for indexing huge quantities of texts remains to be completed.

References

1. Baeza-Yates, R., Gonnet, G.: Fast text searching for regular expressions or automation searching on tries. Journal of ACM 43(6), 915–936 (1996)
2. Blurn, A., Furst, M.: Fast planning through planning graph analysis. Artificial Intelligence 90(1-2), 279–298 (1997)
3. Charikar, M., Indyk, P., Panigrahy, R.: New Algorithms for Subset Query, Partial Match, Orthogonal Range Searching, and Related Problems. In: Widmayer, P., Triguero, F., Morales, R., Hennessy, M., Eidenbenz, S., Conejo, R. (eds.) ICALP 2002. LNCS, vol. 2380, pp. 451–462. Springer, Heidelberg (2002)

4. Doorenbos, R.: Combining left and right unlinking for matching a large number of learned rules. In: AAAI 1994, pp. 451–458 (1994)
5. Flach, P.A., Savnik, I.: Database dependency discovery: a machine learning approach. AI Communications 12(3), 139–160 (1999)
6. Forgy, C., Rete: A fast algorithm for the many pattern/many object pattern match problem. Artificial Intelligence 19, 17–37 (1982)
7. Helmer, S., Moerkotte, G.: A performance study of Four Index Structures for Set-Valued Attributes of Low Cardinality. The VLDB Journal - The International Journal on Very Large Data Bases 12(3), 244–261 (2003)
8. Hoffmann, J., Koehler, J.: A New Method to Index and Query Sets. IJCAI (1999)
9. Mamoulis, N., Cheung, D.W., Lian, W.: Similarity Search in Sets and Categorical Data Using the Signature Tree. In: ICDE (2003)
10. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery Journal 1(3), 241–258 (1997)
11. Manning, C.D., Raghavan, P., Schütze, H.: An Introduction to Information Retrieval, Draft. Cambridge University Press (2009)
12. Melnik, S., Garcia-Molina, H.: Adaptive Algorithms for Set Containment Joins. ACM Transactions on Database Systems 28(2), 1–38 (2003)
13. Rivest, R.: Partial-Match Retrieval Algorithms. SIAM Journal on Computing 5(1) (1976)
14. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. MIT Press (2001)
15. Savnik, I., Flach, P.A.: Bottom-up Induction of Functional Dependencies from Relations. In: Proc. of KDD 1993 Workshop: Knowledge Discovery from Databases, pp. 174–185. AAAI Press, Washington (1993)
16. Savnik, I.: Efficient subset and superset queries. In: Local Proceedings and Materials of Doctoral Consortium of the Tenth International Baltic Conference on Databases and Information Systems (2012)
17. Savnik, I., Flach, P.A.: Discovery of multivalued dependencies from relations. Intelligent Data Analysis Journal 4, 195–211 (2000)
18. Terrovitis, M., Passas, S., Vassiliadis, P., Sellis, T.: A Combination of Trie-trees and Inverted Files for the Indexing of Set-valued Attributes. In: Proc. of ACM International Conference on Information and Knowledge Management (2006)
19. Tousidou, E., Bozanis, P., Manolopoulos, Y.: Signature-based Structures for Objects with Set-valued Attributes. Information Systems 27, 93–121 (2002)
20. Zhang, C., Naughton, J., DeWitt, D., Luo, Q., Lohman, G.: On Supporting Containment Queries in Relational Database Management Systems. ACM SIGMOD (2001)

Opinion Mining in Conversational Content within Web Discussions and Commentaries

Kristína Machová and Lukáš Marhefka

Dept. of Cybernetics and Artificial Intelligence, Technical University, Letná 9, 042 00, Košice, Slovakia
kristina.machova@tuke.sk, lukas.marhefka@student.tuke.sk

Abstract. The paper focuses on the problem of opinion classification related to web discussions and commentaries. It introduces various approaches known in this field. It also describes novelty methods, which have been designed for short conversational content processing with emphasis on dynamic analysis. This dynamic analysis is focused mainly on processing of negations and intensifiers within the opinion analysis. The contribution describes implementations of these methods for the Slovak language. The Slovak dictionaries have been created to support these implementations of lexicon based approach. In addition, the paper describes test results of the presented implementations and discussion of these results as well.

Keywords: Opinion mining, conversational content, opinion classification, dynamic coefficient, n-grams.

1 Introduction

The social web or web with various forms of conversational content increases interactions between users. These interactions are provided in the form of "point-to-point" or "multicast" on-line web services, as chats, discussion forums, IRC (Internet Relay Chat), blog and micro-blog platforms and so on. These services create a big amount of text and therefore they offer interesting possibilities for research in the field of opinion classification.

Where are some interactions, there people influence each other. This influence concerns decision making about various life situations, for example decisions about purchase or production and selling some products, decisions about voting political representatives and so on. This decision making can be supported by information mining from web discussions, reviews, conversations about matter of our interest. Thus, our decision processes can be encouraged by web conversational content. The mentioned conversational content can provide us also with cultural information about films, books and cultural events. On the other hand, it can be a source of some information connected with safety issues, for example suspicious activities or characters, connected with racism, pedophilia or terrorism.

But sometimes these web discussions can be too long and some parts of these discussions can be less informative. That is why there is a need of some tools for

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 149–161, 2013.

automatic analysis of web discussions from the point of positivity or negativity. Our implementation of various methods of opinion analysis provides one such tool.

An opinion represents some positive or negative attitude, view, approach, or emotion of some person – an opinion owner. Opinions are related to a given entity or some of its parts. The entity is some product, person, event, organization, etc. It is the object, we are talking about. It is composed from components. The entity can be some product, for example a mobile. It has some parts – attributes, for example display, design and size, functionalities and so on.

There can be identified a set of different tasks within the field of the opinion analysis:

- *emotion detection* (Does the conversation obtain some emotion?)
- *opinion spam detection* (Is the content of some discussion contribution informative?)
- *subjectivity analysis* (Does the conversation obtain subjective opinion?)
- *opinion polarity classification* (Is the conversation about given subject positive or negative?)
- *theme modeling* (What is the theme/object of this conversation? Is this theme a suspicious one?)
- *authorship identification* (Who is the opinion author? What kind of person is the author of a given contribution?)

Our approach, presented within this paper, is focused on the problem of opinion polarity classification, shortly on opinion classification.

2 Related Works

There are some approaches similar to our approach to opinion classification [2], [11], and [12]. New technical contribution of our approach in comparison with existing works will be discussed within this section. The most similar to our approach is Taboada at al: "Lexicon-Based Methods for Sentiment Analysis" in [11]. It uses a dictionary of words annotated with their orientation (polarity). This approach splits the dictionary into more sub-dictionaries according to word classes (adjectives, nouns, verbs and adverbs). We use only one dictionary (for dynamic coefficient) or two dictionaries (for n-grams). The novelty of our approach is, that the dictionary is generated directly from web discussions, which increases the precision of opinion analysis of the given discussions. In [11], intensification is provided by increasing (respectively decreasing) the semantic intensity of *neighboring* lexical items using a special dictionary of intensifiers. Our approach is more flexible because an intensifier and the related word need not to be neighbors. They can take any position within one lexical unit, while their distance is limited by the given dynamic coefficient. The intensifier can be located before or after the related word. Within our approach, not only intensification, but also negation processing is different, based on processing various combinations of words (lexical units) defined with the aid of dynamic coefficient. All lexical units are sequentially classified into six categories (3, 2, 1, -1, -2, -3). Three of them represents positive polarity and the other three categories negative polarity (with meaning: strong + intensifier, strong or gentle).

Another approach presented in Thelwall at al: "Sentiment strength detection in short informal text" [12] is focused on the SentiStrength detection algorithm, which solves some problems connected with sentiment analysis (generation of the sentiment strength list, optimization of the sentiment word strengths, allocation of the miss words, spelling correction, creation of the booster word list, the negating word list and emoticon list, repeated letters processing and ignoring negative emotion in questions). They devote more effort to correction of non-standard spelling in informal texts. Our approach is not aiming at correct spelling of contributions since the dictionary can easily accommodate some misspelled words as well. Within our approach, the dictionary is generated directly from the analysed discussion and therefore words with sentiment are accepted in spite of common mistakes (for example very common typo using "y" instead of "i" and vice versa in the Slovak language). The algorithm described in [12] was tested on data from MySpace and on wide variety of themes. We have provided our tests on data from narrow domains and therefore our results were a little bit better.

Paper "Learning with compositional semantics as structural inference for substantial sentiment analysis" [2] written by Choi and Cardie is focused on sentiment analysis similarly to our approach. It presents a novel learning based approach that incorporates inference rules inspired by compositional semantics into the learning procedure. Our approach differs from this work because our method is dictionary based (we use simple bag-of-word approach) and it is not machine learning oriented. Our method incorporates surrounding of processed word up to distance K (maximally K neighboring words from the given word). Design of the method in [2] represents meaning of a composed expression as a function of the meanings of its parts within the compositional semantics. In our approach, these parts are lexical units, lengths and number of which are defined by the dynamic coefficient. The approach presented in [2] processes negations and intensification separately. This processing is made over the whole text with the aid of "voting-based inference". In our approach, the negation and intensifications are processed by the same mechanism of using dynamic coefficient. This processing is made separately in each lexical unit, not as the majority vote. Our approach does not use natural language processing. It is based on some statistical principles and so it better processes longer texts, which contain more than one sentence and very short texts (one sentence) can be analyzed with lower precision.

The machine learning approaches (Naïve Bayes classifier, Support Vector Machines (SVM) and Maximal Entropy) were used in [4], which is focused on the automatic classification of opinions from the micro-blog service Twitter. Within [4] the input data for SVM are represented by vectors with dimension **m**. Each item of this vector represents an attribute in the tweet. They used unigram extractor and so each attribute is one word. If this word is presented in the tweet, then its value in the vector is 1. Otherwise its value is 0. This substitution of the frequency of the given word occurrence by simple presence function leads to speedup of a tweet processing.

Sometimes, the introduced opinion analysis is denoted as opinion mining, because it focuses on the extraction of positive or negative attitude of a participant to the commented objects with the aid of mining techniques applied to text documents.

Opinion mining can be extended from the level of whole texts perception to the level of extraction of properties of those objects which match users' interests [3]. Parallel approach to opinion mining is sentiment analysis [10]. Deeper view on sentiment analysis, which is presented in [7], focuses on feature selection. Different approach to web discussion processing is represented by the estimation of authority degree of some information sources, for example of actors contributing to discussion forums or social networkss. An important technique for authoritative actors searching is visualization approach, which is introduced in [5]. Some effort was spent on semantically enriching algorithms for analysis of web discussion contributions by authors of [8]. Also dedicated information can be used as an interface to newsgroup discussions [9].

Nowadays, opinion analysis has become an important part of social networks analysis. Existing opinion analysis systems use large vocabularies for opinion classification into positive or negative answer categories. Such approach was used in [1]. Authors studied accuracy of the opinion analysis of Spanish documents originated in the field of economic. This approach uses a regression model for classification into negative or positive opinions. Authors studied how quality depends on the granularity of opinions and rules, which were used in the regression model. Another study [6] was focused on the possibility of using lesser granularity without any significant precision decrease. The results of this study show no substantial difference between one and two parameter regression models as well as no statistically significant difference between models with different granularity. Thus, for example, simpler models can be used with the sentiment scale reduced to five degrees only. The approach, presented in this paper, uses a scale with five degrees for opinion classification as well, but it differs from the previous approaches in vocabulary cardinality. Our work focuses on creating vocabularies with strong orientation on the discussion domain, not so large but created directly from live discussions. We do not use regression models.

3 Opinion Mining

The process of the opinion classification consists of minimally two subtasks:

1. *input text preprocessing* or transformation of discussion texts into lexical units which can be easily processed,
2. *opinion classification*, which represents determination (calculation) of the polarity (positive or negative) of a subjective opinion.

3.1 Input Text Preprocessing

There are several approaches how to parse an analyzed input text into smaller lexical units. The most used approaches are n-grams and part-of-speech tagging.

N-gram can be defined as a series of items from some sequence. From the semantic point of view, it can be a sequence of phones, characters or words. In practice, n-gram as a sequence of words is the most common. The sequence of two (three) words is called bigram (trigram). For the case, when more than three words (exactly n) are in

the same sequence, such sequence is called n-gram. N-grams are used in the wide scale of fields, as theoretical mathematics, biology, cartography, even in the field of music.

In the field of natural language processing, n-grams can be used for words prediction. The words prediction uses so called "n-grams model". This n-gram model calculates the probability of occurrence of the last word in an n-gram from the previous n-grams. Another way of using n-grams is plagiarism discovery by text dividing into smaller fragments. These fragments are represented by n-grams. These n-grams can be easily compared and consequently, the measure of similarity of comparing documents can be calculated. N-grams are often used for text categorization and also for effective searching for correct candidates of misspelled words. Our approach, presented within this contribution uses n-grams for splitting the web discussion contributions into lexical units.

The part-of-speech (POS) tagging represents the recognition of word class within the text on the basis of the given language attributes and relationships between word classes. The POS tagger is a program, which is able to read text and to assign word class to each word from the given text. Such word class assigning can be based on the word definition or the word position within the sentence. There are three types of POS taggers: rule taggers, stochastic taggers and transformation taggers. The rule tagging algorithms are based on sets of rules, which are used for tagging of the processed text. This tagged text can be used as a training set for stochastic approach. The stochastic taggers are based on probability of the given tag occurrence within the given text. The stochastic approach requires an input training set. The last one, transformation taggers represent combination of two previous approaches.

3.2 Opinion Classification

There are two basic approaches to opinion classification [11]:

- *machine learning approach* (classification based approach), which uses the statistical and machine learning methods
- *lexicon based approach*, which can be based on the dictionary (*dictionary based approach*) or corpus.

Machine Learning Approach. This approach is based on using some of well-known methods of machine learning, for example Naïve Bayes classifier, Support Vector Machines (SVM) and Maximal Entropy and so on.

Naïve Bayes is a simple classifier based on Bayes formula of conditional probability with the big degree of freedom. A problem is defined as a classification of a given tweet **d** to a class **c**. This class **c** represents positive or negative polarity of some contribution. The original Bayes theorem in the formula (1)

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \tag{1}$$

was modified taking into account the following fact: the probability of a tweet occurrence is 100% and so probability p(d|c) was replaced by probability of feature $\mathbf{f_i}$ in the class **c**. The final Bayes formula for conversational content is following (2)

$$P_{NB}(c \mid d) = \frac{p(c)\prod_{i=1}^{m} p(f_i \mid c)^{n_i(d)}}{p(d)} . \quad (2)$$

Maximal Entropy. Within the information theory, the entropy represents a measure of indeterminism. The principle of maximal entropy is an axiom of the Bayes probability theory. It states that the probability distribution, which can represent current state of knowledge in the best way, is the distribution with the highest information entropy. This model is described by the following formula (3)

$$P_{ME}(c \mid d, \lambda) = \frac{\exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right)}{\sum_{c'} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right)} . \quad (3)$$

Within this formula: **c** is positive (negative) class, **d** is a tweet and λ is a weighting vector. The weighting vector represents an importance of the given feature **F** for the given class assignment (usually the word with strong positive meaning). The weighting vector is set on the base of numerical optimization of its components by the maximization of conditional probabilities. The classification was provided by Standford classifier [13].

Support Vector Machines (SVM) is the groups of machine learning classification methods. The difference between the SVM and Perceptron is that SVM can be also used for nonlinear classification.

Dictionary Based Approach. This approach looks on the input text as a set of words. It does not take into account relationships between words within sentences or grammatical rules. Not every word has the same importance in the opinion classification. It is important to find within the processed text mainly words, which can express opinion in the best way. Such words are stored in the classification dictionary. The dictionary has to obtain some words with sentiment before starting the process of opinion classification. On the base of known polarity of such characteristic words, polarity of whole texts from conversation and finally polarity of the whole web discussion can be determined.

The classification dictionary is a database of words. These words are descriptors with marked influence on polarity determination, mainly adjectives, adverbs, nouns and verbs. The simplest dictionaries are the dictionaries which enable binary classification. More suitable are fuzzy dictionaries, which are able to determine not only polarity but also the strength of the polarity. Such dictionary can be created for only a given application or some known lexicons can be used, for example Word Net, WordNet-Affect, SenticNet, SentiWordNet and so on.

Each dictionary should contain words, which are common within web conversation forums, blogs, commentaries and so on including slang words, words without diacritics and words with more common grammatical mistakes. Our approach, presented within this contribution is the dictionary based approach.

4 Negation and Intensification

Besides basic problems of opinion analysis as word subjectivity identification, word polarity (orientation) determination and determination of intensity of the polarity, each application of opinion classification has to solve the processing of negation and intensification. The basic problems can be simply solved with the aid of classification dictionaries. These dictionaries focus on those words, which are able to express subjectivity very well - mainly adjectives (e.g. 'extraordinary') and adverbs (e.g. 'awfully') are considered. On the other hand, other word classes must be considered as well in order to achieve satisfactory precision, for example nouns (e.g. 'crash') or verbs (e.g. 'damage'). The words with subjectivity are important for opinion classification. Therefore, they are identified and inserted into one vocabulary or into a set of vocabularies – one vocabulary for each word class (adjectives dictionary, adverbs dictionary, nouns dictionary and verbs dictionary). Words with subjectivity are inserted into the corresponding vocabulary together with their degree of polarity. The majority of opinion classification applications work with 5 to 10 degrees of polarity.

4.1 Negation Processing

There are many approaches to negation processing:

- Switch negation,
- Shift negation,
- Dynamic coefficient usage.

The *switch negation* is simply reversion of the polarity of the lexical item. The reversion represents changing the number sign of the polarity degree (from minus to plus and vice versa). There are many various words related to negation that need to be taken into account: *not, none, never, nothing*, which usually are situated next to a related word – item. But also other negations should be taken into account as *without, don't, lack* and so on, which can be situated in significant distance from the lexical item. These negations can be hardly processed by switch negation. What is more, switch negation can be insufficiently precise, because negation of a strong positive word rarely is a strong negative word and vice versa. More often the negation of a strong positive word is a slightly negative word and vice versa.

The *shift negation* focuses on the case, when negation of a strong positive (negative) word is not strong negative (positive). Instead of changing the sign, "shift negation" shifts polarity degree toward the opposite polarity by a fixed value. For example, if the shift value is "6", then negation can be calculated in the following way: "It is not completely dysfunctional ($-5 + 6 = +1$) but also not very useful ($+5 - 6 = -1$)."

The *dynamic coefficient* usage does not need to suppose that all negation words are situated next to related words – items. In this method, the text of conversation is divided into lexical units of the length given by dynamic coefficient. It does not matter how far the negation is from the related word within the same lexical unit. One sentence can be transformed into one or more lexical units.

4.2 Intensification

There are two different approaches to intensification:

- Dictionary based intensification,
- Intensification based on dynamic coefficient.

The *dictionary based intensification* supposes the dictionary of intensifications – words, which are able to increase (or decrease) the intensity of polarity. Each of intensifications should be stored in dictionary together with sign and number. This number can represent a percentage of changing of polarity intensity and sign represents the type of this change. This approach supposes that a lexical item is situated next to intensifier changing the polarity degree. This condition is fulfilled only in small number of cases.

The *intensification based on dynamic coefficient* does not suppose that a lexical item is situated next to intensifier changing the polarity degree. This kind of intensification does not depend on how far the negation is from the related word within the same lexical unit.

Our implementation uses dynamic coefficient for negation processing and also for intensification.

5 Dynamic Coefficient in Opinion Classification

First, a simple version of Opinion Classification Application (OCA) has been designed and implemented. This OCA represented the dictionary based approach to opinion classification with "*static coefficient*", which was defined by user as the parameter of the application. This static coefficient designates the length of a lexical unit (number of words from the text, which belong to one lexical unit) for processing within texts - contributions of discussion forums. This processing represents the classification to positive or negative opinions. The OCA solves basic problems of opinion classification (words with subjectivity identification, words polarity determination and polarity intensity determination) but also problems connected with word polarity reversion by negation and intensification of words. The solving of these problems desire to process a combination of words instead of separated words. A length of processed word combination can be represented by the static coefficient, but also by a dynamic coefficient.

The value of the "*dynamic coefficient*" *K* is being dynamically changed during processing of different lexical units. The dynamic coefficient adapts itself to the length of a lexical unit (sequence of words) under investigation. The value *K* represents the number of words, which are included into the same word combination. In the case, when the value is higher than the number of words in the sentence, this value is dynamically decreased in order to ensure, that the combination contains only words from the investigated sentence, not from the beginning of the next sentence.

Three ways of the dynamic coefficient setting have been proposed. Before the process of opinion classification can start, user has to select one of the methods for coefficient determination, as it is illustrated in Fig 1.

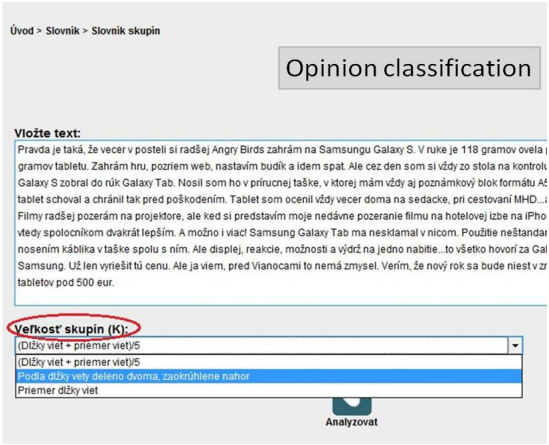


Fig. 1. Three ways of the dynamic coefficient determination

The first used way of dynamic coefficient setting is the calculation of the *average length of all sentences* of the discussion contribution text, which is analyzed. So, each contribution text from a live discussion can have the different value of the dynamic coefficient. The calculated value of the dynamic coefficient is used for processing of all sentences of the given contribution. Thus, each sentence of the same contribution is processed using the same value of the dynamic coefficient, although actual length of these sentences is different.

The second used way of dynamic coefficient setting is the calculation of the *half length of each sentence* of the discussion contribution text. If needed, the calculated value is rounded up. Each analyzed sentence is processed using a different value of the dynamic coefficient.

The last used way is the *hybrid approach***,** which determines the value of the dynamic coefficient as an average of two values obtained from *Average length of all sentences* method and *Half length of each sentence* method, respectively.

6 N-grams Application to Opinion Classification

Our n-grams application to opinion classification belongs to dictionary approach. This application is oriented on the Slovak language, so it uses a dictionary of Slovak language words. The structure of this dictionary is given by four more important word classes for opinion analysis: adjectives, adverbs, nouns and verbs. The dictionary consists of two different parts. The first part contains adjectives, nouns and verbs. The second one contains adverbs and negations. The first part of the dictionary is used for solving basic problems of opinion classification. The second part of the dictionary is used for negation processing and intensification because the adverbs have in a language the function to increase (*"surprisingly nice"*) or to decrease (*"extremely low-class"*) intensity of a word polarity. The dictionary contains also some emoticons, which naturally can express emotions and opinions very well. Sometimes, the text analysis can be less clear and so emoticons can increase the precision of the classification. These emoticons are stored within the first part of the dictionary.

All words and emoticons from this first dictionary are quantified to polarity degree within the interval from -3 to 3. With respect to the second part of dictionary, intensifiers (adverbs) are assigned by value from -0.5 to 1 and negations are represented by value -2. The analyzed text is processed by the following way. All words from a text are compared with all words in the first dictionary, and in the case of match, the value of the word found in the dictionary is added into the overall sum of values of analyzed words. This sum is consequently multiplied by the second sum of values obtained from the second dictionary (for intensifiers and negations). This processing can be represented by formula (4)

$$P = \sum v(w_i^1)[1+\sum v(w_j^2)] \; . \tag{4}$$

Within this formula:

P …is the polarity degree of analyzed text
$v(w_i^1)$ …is the value of word w_i of text found in the first part of the dictionary
$v(w_j^2)$ …is the value of word w_j of text found in the second part of the dictionary.

For example:

- the lexical unit *"The chair was comfortable but dirty and its color was awful.*" is processed by the following way:
  comfortable (+1) + dirty (-3) + awful (-1): P = -3

- the lexical unit *"It is not good idea.*" is processed by the following way:
  good (+1) first dictionary, not (-2) second dictionary: P = 1*[1+(-2)] = -1

- the lexical unit *"The text processing is very decent.*" is processed by the following way: decent (+1) first dictionary, very (+1) second dictionary:
  P = 1*[1+1)] = +2.

The question is, how long the length of processed lexical unit should be. We decided to represent the lexical units by n-grams. The value of *n* was fixed on 4 in an experimental way. Shorter value of *n* can separate negation or intensifier from related word. Consequently, we have designed new versions of n-gram approach. The first version has created 4-grams applying shift by one position. Another one was based on comparison of polarity value obtained by 4-grams and polarity value obtained by1-grams.

7 Evaluation and Testing

The implementations of various versions of opinion classification were tested within several experiments.

First, the basic version with static coefficient (called "Static coefficient" in Table 1) was tested on the set of discussion contributions from the portal http://www.mobilmania.sk (a discussion thread related to reviews of the LGKU990 mobile telephone). This set of contributions contained 1558 words within 236 lexical units and the used classification dictionary had 27 positive words, 27 negative words,

10 negations and 11 intensifications. The resulting precision of these tests was 0.78, which was an average calculated from precision of positive contributions classification (higher value 0.86) and precision of negative ones (lower value 0.69).

The next versions were the version using the dynamic coefficient, which was set as an average length of each sentence (called "Dynamic coefficient 1" in the Table 1), the version using the dynamic coefficient determined as half length of each sentence (called "Dynamic coefficient 2" in the Table 1) and the version using the hybrid approach (called "Hybrid" in the Table 1) were tested on 50 reviews (25 positive reviews and 25 negative ones) obtained from the page http://recenzie.sme.sk. The classification dictionary contained 150 words in all categories (positive words, negative words, negations and intensifications). The results of negative contributions classification were surprisingly high as can be seen in Table 1 (0.84, 0.88 and 0.84) – higher than precision of positive contributions classification (0.76, 0.80 and 0.80). It may be caused by the fact, that the number of negative contributions was the same as the number of positive ones. Usually, the precision of negative contributions is lower because the number of available negative contributions is fairly lower.

Lastly, the version based on n-grams was tested in two independent experiments (called "n-grams 1 and n-grams 2" in Table 1). The first experiment was performed on a set of 42 contributions with 2350 words. These contributions were extracted from discussion forums available on http://www.mojandroid.sk (a discussion thread related to reviews of the mobile telephones HTC One X and HCT One S) and http://www.pocitace.sme.sk (a discussion thread related to reviews of two products Asus Transformer Prime TF201 and Asus Transformer Pad TF300T). The second experiment (n-grams 2) was performed on a set of 71 contributions (4341 words) from the portal http://tech.sme.sk (a discussion thread related to reviews of the telephone Samsung Galaxy S4) and from the portal http://www.mojandroid.sk (a discussion thread related to reviews of the telephones HTC ONE and Samsung Galaxy S4). These two tests of n-grams application were provided using classification dictionary with 44 positive words, 46 negative words, 3 kinds of negations and 10 intensifications. The achieved precision of classification of positive contributions in the first experiment (N-grams 1) was quite satisfactory (0.83) but in the second experiment (N-grams 2) was not very high (0.76). A different situation was in the case of negative contributions classification. The achieved precision in the first experiment was only 0.57 and in the second experiment even worse 0.42, what is very low value. Maybe, an extremely low number of available negative contributions within these two last experiments was the cause of such disappointed results.

Table 1. Precision of testing results of various implementation versions

Version	Positive	Negative	Average precision
Static coefficient	0.86	0.69	0.78
Dynamic coefficient 1	0.76	0.84	0.80
Dynamic coefficient 2	0.80	0.88	0.84
Hybrid	0.80	0.84	0.82
N-grams 1	0.83	0.57	0.70
N-grams 2	0.76	0.42	0.59

The best average precision was achieved by three implementation versions using the dynamic coefficient. These results are interesting, because these versions had better results for processing negative contributions than for processing positive contributions. This fact is not common within the existing opinion classification applications. Usually, a processing of negative contributions is less precise. The basic version using static coefficient and version based on n-grams are more common from this point of view, because they achieved lower precision within negative contributions while they had quite high precision within positive contributions processing. The higher value of precision than 0.80 (80 percentage) seems to be very good in comparison with the average precision of human advices (about 70 percentage). An important factor is also the language which is used to present the classified opinions. Although languages have some similar features, the complexity of the task is evaluated on the basis of the emergent expression of the given particular language.

Our modified application achieves relatively better results - higher precision (0.84, 0.82 using dynamic coefficient) against [11]. The term "relatively" is used because two kinds of testing results achieved on two different input data sets were compared. They present many tests with resulting performance up to 80 percentages (within more than 100 experiments only two results are higher than 80 percentage). On the other hand, our tests were not so complex as in [11] and therefore our resulting precision could be lower when using their corpuses of reviews. Our results seem to be better (once again relatively) than results (up to 73 percentages) in [12]. This comparison is relative because of compared results were achieved on different input contributions and because no cross validation was used. On the other hand, performance achieved in [2] is higher than performance of our applications.

8 Conclusions

The paper introduced variety of approaches to solving the problem of opinion classification to positive or negative polarity. It also described five various original methods of opinion classifications and it also introduced the test results of implementations of the presented methods.

The novelty of our approach is, that our dictionary was generated directly from a web discussion, which increases the precision of opinion analysis of the given discussion and so the dictionary can easily accommodate some misspelled words as well.

The achieved precision of classification of positive contributions within all presented versions of our application was approximately 80 percentages. It is not so bad in the comparison with other existing applications. On the other hand, precision of classification of negative contributions using n-grams is too low. For the future, we would like to test the introduced version of opinion classification application on larger corpus of discussion contributions.

There is a possibility to upgrade version based on n-grams to achieve higher precision within processing of negative contributions. There is a need to extend the classification dictionary. This version should be enriched by techniques for processing also contributions, which contain only neutral words, but their context is positive or negative. The techniques for processing of irony and ambiguity should be included too. The research in the field of opinion classification has big importance for the

future. A successful application of opinion classification can be very helpful in the process of decision making.

Acknowledgements. The work presented in this paper was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic within the 1/1147/12 project "Methods for analysis of collaborative processes mediated by information systems".

References

1. Catena, A., Alexandrov, M., Ponomareva, N.: Opinion Analysis of Publications on Economics with a Limited Vocabulary of Sentiments. International Journal on Social Media - MMM: Monitoring, Measurement, and Mining 1(1), 20–31 (2010)
2. Choi, Y., Cardie, C.: Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In: Proc. of the EMNLP 2008, Conference on Empirical Methods in Natural Language Processing, pp. 793–801 (2008)
3. Ding, X., Liu, B., YuA, P.: Holistic Lexicon-Based Approach to Opinion Mining. In: Proc. of the Int. Conf. on Web Search and Web Data Mining WSDM 2008, New York, NY, USA, pp. 231–240 (2008)
4. Go, A.: Twitter Sentiment Classification using Distant Supervision. Stanford University, http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf
5. Heer, J., Boyd, D.: Vizster: Visualizing Online Social Networks. In: Proceedings of the IEEE Symposium on Information Visualization INFOVIS 2005, Washington, USA, pp. 5–13 (2005)
6. Kaurova, O., Alexandrov, M., Ponomareva, N.: The Study of Sentiment Word Granularity for Opinion Analysis (a Comparison with Maite Taboada Works). International Journal on Social Media - MMM: Monitoring, Measurement, and Mining 1(1), 45–57 (2010)
7. Koncz, P., Paralič, J.: An Approach to Feature Selection for Sentiment Analysis. In: Proc. of the INES 2011 - 15th International Conference on Intelligent Engineering Systems, Poprad, pp. 357–362 (2011) ISBN 978-142448956-5
8. Lukáč, G., Butka, P., Mach, M.: Semantically-enhanced Extension of the Discussion Analysis Algorithm in SAKE. In: SAMI 2008, 6th International Symposium on Applied Machine Intelligence and Informatics, Herľany, Slovakia, pp. 241–246 (January 2008)
9. Mach, M., Lukáč, G.: A Dedicated Information Collection as an Interface to Newsgroup Discussions. In: IIS 2007 - 18th International Conference on Information and Intelligent Systems, Varazdin, Croatia, September 12-14, pp. 163–169 (2007) ISBN 978-953-6071-30-2
10. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval 2(1-2), 1–135 (2008)
11. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37(2), 267–307 (2011)
12. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science and Technology 61(12), 2544–2558 (2010)
13. Stanford Classifier. Stanford University, http://nlp.stanford.edu/software/classifier.shtml

Diagnosis of Higher-Order Discrete-Event Systems

Gianfranco Lamperti[1] and Xiangfu Zhao[2]

[1] Dipartimento di Ingegneria dell'Informazione,
Università degli Studi di Brescia, Italy
[2] College of Mathematics, Physics and Information Engineering,
Zhejiang Normal University, China

Abstract. Preventing major events, like the India blackout in 2012 or the Fukushima nuclear disaster in 2011, is vital for the safety of society. Automated diagnosis may play an important role in this prevention. However, a gap still exists between the complexity of systems such these and the effectiveness of state-of-the-art diagnosis techniques. The contribution of this paper is twofold: the definition of a novel class of discrete-event systems (DESs), called higher-order DESs (HDESs), and the formalization of a relevant diagnosis technique. HDESs are structured hierarchically in several cohabiting subsystems, accommodated at different abstraction levels, each one living its own life, as happens in living beings. The communication between subsystems at different levels relies on complex events, occurring when specific patterns of transitions are matched. Diagnosis of HDESs is scalable, context-sensitive, and in a way intelligent.

1 Introduction

In the last decades, automated diagnosis of complex systems has become increasingly important for the safety of society. It suffices to consider two recent fateful events: the 2012 India blackout, and the 2011 Fukushima Daiichi nuclear disaster.

In July 2012, India suffered from a major blackout, the largest power outage in history, occurring as two separate events (on 30 and 31 July), which affected over 620 million people (half of India's population), and spread across 22 states, with an estimated 32 gigawatts of generating capacity being taken offline.

Among other consequences, the outage caused chaos in rush hours, as passenger trains were shut down and traffic signals were non-operational. Several hospitals reported interruptions in health services. Hundreds of miners were trapped underground due to failures in lifts. Water treatment points were shut down for hours and millions of people were not able to draw water from wells powered by electric pumps.

In August 2012, the investigation committee concluded that among other factors responsible for the blackout was the loss of a 400V transmission line caused by *misbehavior of the protection system* [1]. The committee also provided several recommendations to prevent further blackouts, including an *audit of the protection system*. Also some technology sources and the United States Agency for International Development (USAID) proposed that another widespread outage could be prevented by an integrated network of microgrids and distributed generation connected seamlessly with the main grid via a superior smart grid technology which includes *automated fault detection*, islanding and self-healing of the network.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 162–177, 2013.

In March 2011, following the Tohoku earthquake and tsunami, Japan was struck by a nuclear disaster caused by a series of *equipment failures*, nuclear meltdowns, and releases of radioactive materials at the Fukushima I Nuclear Power Plant. It was the largest nuclear disaster since Chernobyl in 1986.

Immediately after the earthquake, the three reactors that were operating in the power plant shut down automatically, and emergency generators were started for controlling electronic devices and coolant systems. However, the tsunami following by the earthquake quickly flooded the underground rooms housing the emergency generators, causing the latter to fail, thereby interrupting the power to the pumps aimed at continuously circulating coolant water to prevent the nuclear reactor from melting down. As a consequence, the reactors overheated owing to the high radioactive decay heat that continued for hours (even days) after the shutdown of the nuclear reactor.

In that situation, what could have prevented the meltdown was prompt flooding of the reactors with sea water. However, since salt in water was bound to ruin the (costly) reactors permanently, this action was delayed and taken only after an explicit order of the government. But it was too late. In the following hours and days all three reactors experienced full meltdown.

Subsequent explosions and the atmospheric venting of radioactive gases led to a 20 Km-radius evacuation around the plant. Sea water exposed to melting rods was returned to the sea heated and radioactive in large volumes for several months. The accident was eventually assessed at Level 7 (the maximum scale value).

Conclusions based on the analysis of the accident procedure manuals used for Fukushima Daiichi nuclear power plant published by the japanese Nuclear and Industry Safety Agency (NISA), include that after the batteries and power supply boards were flooded, almost all electricity sources were lost, and this event was not envisioned. Besides, a number of nuclear energy specialists remarked that plants should be able to *maintain electricity* during an earthquake (and other emergency situations).

The investigation report released by a government-appointed panel held in June 2012 asserts that the failure in preventing the nuclear disaster was not caused by the fact that a large tsunami was unanticipated, but because of the reluctance in investing time, effort, and money in protecting against a natural disaster considered unlikely [24].

The two disasters in India and Japan share two common properties:

- *Complexity* of the monitored system: in both cases, the system is composed of several interconnected subsystems, possibly at different abstraction levels;
- *Dependency* on electricity: in both cases, the failure in supplying electricity plays a major role in provoking the disaster.

A complex system is not necessarily large (even though a large system is likely to be complex). In our meaning, complexity refers to the mode in which the system is organized, at different levels of abstraction, with each level being characterized by its proper behavior, which depends on the behaviors of lower-level layers, yet differs from just the composition of them. We call it *behavior stratification*.

A large system that is composed of many interacting components at the same level of abstraction is not complex *per se*. The human brain is complex not merely because it is composed of billions of neurons but because this huge biological neural network is organized in a hierarchy of different layered brains: the primitive reptilian brain at

the bottom, the emotional mammal brain in the middle, and the rational primate brain on top, with each brain being composed of several parts (amygdala, prefrontal cortex, temporal lobes, *etc*.).

In this paper, we apply behavior stratification to discrete-event systems (DESs) [5], in particular, to a class of asynchronous DESs called active systems [16]. DESs are typically modeled as networks of interacting components, where the behavior of each component is described by a communicating automaton [3]. However, complexity of the DES has become a research issue only recently. Previous research has mainly focused on relevant yet different aspects, including incrementality [2, 9], distribution / decentralization [20, 6, 22, 7, 8, 21, 23], and uncertainty / incompleteness [15, 25, 18, 14, 26].

The notion of context-sensitive diagnosis was introduced for DESs that are organized within abstraction hierarchies, so that candidate diagnoses can be generated at different abstraction levels [17]. Even in that work, albeit the diagnosis depends on the context, the DES is assumed to be a network of components without behavior stratification. When behavior stratification occurs, we have a *higher-order DES* (HDES).

2 Higher-Order Discrete-Event Systems

A higher-order DES, namely $\mathcal{H}$, is a tree where nodes are *components*. Leaf nodes are *basic components*, while internal nodes are *complex components*. The set of child components of a complex component X is indicated by $\mathcal{C}(X)$. Each (either basic or complex) component in $\mathcal{H}$ is defined in terms of a *topological model* and a *behavioral model*. The topological model consists of a set of *input terminals* and a set of *output terminals*. Components in $\mathcal{C}(X)$ are connected to one another through *links*, with each link exiting the output terminal of one component and entering the input terminal of another component. These connections form a network $\mathcal{N}(X)$.

Let $\mathbf{I}$ and $\mathbf{O}$ denote the input and output terminals of a component C. The behavioral model of C is a communicating automaton $(\mathbb{S}, \mathbb{I}, \mathbb{O}, \mathbb{T})$, where $\mathbb{S}$ is the set of states, $\mathbb{I}$ the set of input events, $\mathbb{O}$ the set of output events, and $\mathbb{T} : \mathbb{S} \times (\mathbb{I} \times \mathbf{I}) \times 2^{(\mathbb{O} \times \mathbf{O})} \mapsto 2^{\mathbb{S}}$ the (nondeterministic) transition function. A transition is triggered by an input event and generates a (possibly empty) set of output events. The latters are thus made available as input events at the corresponding input terminals of connected components, while the input (triggering) event is consumed. A transition can be triggered only if all links, towards which output events are generated, are empty (no event is in the link).

Each complex component X is endowed with a *Cot* additional input terminal, which is sensitive to *complex events*. A complex event is a set of *pattern events*. A pattern event occurs when the network $\mathcal{N}(X)$ undergoes a string of transitions matching a given regular expression. The alphabet of such a regular expression is the whole set of transitions of components in $\mathcal{C}(X)$.

In general, for each complex component X, a set $\mathcal{P}(X)$ of *patterns* is defined, with each pattern being a pair (p, r), where p is the name of a pattern event and r a regular expression on transitions of $\mathcal{C}(X)$. Several pattern events may occur simultaneously. In fact, given a string $\mathcal{T}$ of transitions of components in $\mathcal{C}(X)$, each suffix of $\mathcal{T}$ matching a regular expression in $\mathcal{P}(X)$ gives rise to a pattern event. The whole set of these pattern events forms a complex event.

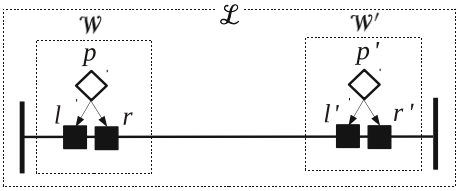


Fig. 1. HDES: protected power transmission line

Example 1. Shown in Fig. 1 is a HDES representing a power transmission line $\mathcal{L}$. On both sides, the line is protected from short circuits by a protection hardware, namely $\mathcal{W}$ and $\mathcal{W}'$. Each of them is composed of a protection device, p and p', respectively, and two breakers, l and r, and l' and r', respectively. Boxes denote complex components $\mathcal{L}$, $\mathcal{W}$, and $\mathcal{W}'$. We assume that the output terminal of the protection device is exited by two links directed to the input terminals of the two breakers. The protection device is sensitive to short circuits on the line, detected as lowering of voltage, in which case it commands the breakers to open in order to isolate the line (just one open breaker on both sides is sufficient for isolation). Once the line is isolated, the short circuit is expected to die. If so, the protection device commands both breakers to close in order to reconnect the line (all breakers need to be closed). However, faulty behavior may occur, specifically:

- The protection device sends the wrong command;
- The breaker does not react to the command of protection device (thereby remaining in its state);
- The protection hardware fails to either disconnect or connect the line (if just one breaker does not open, the protection hardware is normal, as disconnection occurs; instead, for the connection, both breakers must close);
- The line is not isolated, or once the short circuit is dead, the line is not reconnected, or after reconnection the short circuit is still alive (e.g. a tree fallen on the line).

Outlined in Fig. 2 are the behavioral models. Details on component transitions are provided in Table 1. Patterns relevant to complex events in Table 1 are defined in Table 2, where '?' means optionality, '$*$' means repetition zero or more times, '$+$' means repetition one or more times, and '$\neg$' (negation) means any transition different from its argument. For instance, consider pattern event *ps* (persistent short circuit) and the corresponding regular expression, which occurs when either $\mathcal{W}$ or $\mathcal{W}'$ repeats one or more times the following sequence of transitions: it closes (w_5) and then, after zero or more occurrences of w_1, it opens again (w_2), followed by zero or more transitions other than w_5. Notice that *ps* is a single pattern event, while **ps** is the singleton $\{ps\}$ (complex event). Incidentally, in our example all complex events are singletons. ◇

2.1 Pattern Space

In order to detect complex events, the state of the matching of patterns is to be maintained somewhere. To this end:

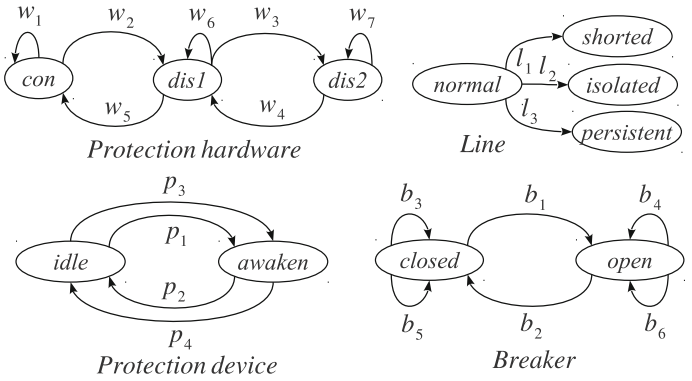


Fig. 2. Behavioral models

Table 1. Details for transitions of behavioral models in Fig. 2

T	*Action performed by component transition* T
p_1	Detects low voltage and outputs *op* (open) event
p_2	Detects normal voltage and outputs *cl* (close) event
p_3	Detects low voltage, yet outputs *cl* event
p_4	Detects normal voltage, yet outputs *op* event
b_1	Consumes *op* event and opens
b_2	Consumes *cl* event and closes
b_3	Consumes *op* event, yet keeps closed
b_4	Consumes *cl* event, yet keeps open
b_5	Consumes *cl* event
b_6	Consumes *op* event
w_1	Consumes **nd** (not disconnected) complex event
w_2, w_3	Consumes **di** (disconnected) complex event
w_4, w_5	Consumes **co** (connected) complex event
w_6, w_7	Consumes **nc** (not connected) complex event
l_1	Consumes **ni** (not isolated) complex event
l_2	Consumes **nr** (not reconnected) complex event
l_3	Consumes **ps** (persistent short) complex event

- For each pattern (p, r), a deterministic *pattern automaton* A equivalent to regular expression r is generated, where final states are marked by pattern event p.
- For each complex component X for which the set $\{A_1, \ldots, A_k\}$ of pattern automata were generated, a *pattern space*, written $Pts(X)$, is created as follows:
  1. A nondeterministic automaton $\mathcal{N}$ is created by generating its initial state S_0 and one empty transition from S_0 to each initial state of A_i, $i \in [1 \,..\, k]$;

2. In each A_i, $i \in [1\,..\,k]$, an empty transition from each non-initial state to S_0 is inserted;[1]
3. $\mathcal{N}$ is determinized into $Pts(X)$, where each final state S is marked by the union **p** of the pattern events that are associated with the states in S that are final in the corresponding pattern automaton.[2]

Table 2. Specification of patterns by regular expressions

Pattern event	*Meaning*	*Regular expression*
di	Disconnected	$b_1(l) \mid b_1(r)$
co	Connected	$b_2(l) \mid b_2(r)$
nd	Not disconnected	$p_3(p) \mid p_1(p)((b_3(l)b_3(r)) \mid (b_3(r)b_3(l)))$
nc	Not connected	$p_4(p) \mid p_2(p)(b_5(r)?\, b_4(l) \mid b_5(l)?\, b_4(r))$
ni	Not isolated	$w_1(\mathcal{W}) \mid w_1(\mathcal{W}')$
nr	Not reconnected	$w_6(\mathcal{W}) \mid w_7(\mathcal{W}) \mid w_6(\mathcal{W}') \mid w_7(\mathcal{W}')$
ps	Persistent short	$(w_5(\mathcal{W})w_1(\mathcal{W})^*w_2(\mathcal{W})(\neg w_5(\mathcal{W}))^*)^+ \mid$ $(w_5(\mathcal{W}')w_1(\mathcal{W}')^*w_2(\mathcal{W}')(\neg w_5(\mathcal{W}'))^*)^+$

Example 2. With reference to Example 1, consider complex component $\mathcal{W}$, whose patterns are defined on top of Table 2.

Following the steps specified above, $Pts(\mathcal{W})$ is generated as detailed in Table 3. The main part of the table represents the transition function, where for each component transition $T \in \{p_1(p), \ldots, b_5(r)\}$ (listed in the first column), and for each state $\mathcal{P}_i$, $i \in [0\,..\,10]$ (listed in the first row), the reached state is indicated in the cell $(T, \mathcal{P}_i)$. Moreover, highlighted states are final, namely $\mathcal{P}_1$, $\mathcal{P}_4$, $\mathcal{P}_7$, and $\mathcal{P}_8$. Complex events associated with final states are listed in the last row, namely **di**, **co**, **nc**, and **nd** (details are in Table 1). These are singletons of the homonymous pattern event. ◇

Proposition 1. *The set* **p** *marking a final state* S_f *of* $Pts(X)$ *is composed of the pattern events* p *such that* $(p, r) \in \mathcal{P}(X)$, $\mathcal{T}$ *is a string in the language of* $Pts(X)$ *ending at* S_f, $\mathcal{T}'$ *is a string matching regular expression* r, *and* $\mathcal{T}'$ *is a suffix of* $\mathcal{T}$.

Proof. (*Sketch*) In creating $Pts(X)$, if we omit step 2 then the language of $Pts(X)$ will be the union of the languages of regular expressions involved in $\mathcal{P}(X)$, where each string ending at S_f matches the regular expression associated with a pattern event in **p**. Consequently, the statement of the theorem should be restricted in the last condition by $\mathcal{T}' = \mathcal{T}$. The more relaxed condition, namely $\mathcal{T}'$ being a suffix of $\mathcal{T}$, comes from step 2 of the construction: generally speaking, because of additional empty transitions, each string $\mathcal{T}$ ending at S_f matches regular expressions only in its suffixes. □

[1] This allows for pattern-matching of overlapping strings.

[2] Each state S of the deterministic automaton is identified by a subset of the states of the equivalent nondeterministic automaton.

Table 3. Tabular specification of pattern space $Pts(\mathcal{W})$

$T \setminus \mathcal{P}_i$	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$	$\mathcal{P}_6$	$\mathcal{P}_7$	$\mathcal{P}_8$	$\mathcal{P}_9$	$\mathcal{P}_{10}$
$p_1(p)$	$\mathcal{P}_2$	$\mathcal{P}_2$	$\mathcal{P}_2$	$\mathcal{P}_2$	$\mathcal{P}_2$	$\mathcal{P}_2$	$\mathcal{P}_2$	$\mathcal{P}_2$	$\mathcal{P}_2$	$\mathcal{P}_2$	$\mathcal{P}_2$
$p_2(p)$	$\mathcal{P}_3$	$\mathcal{P}_3$	$\mathcal{P}_3$	$\mathcal{P}_3$	$\mathcal{P}_3$	$\mathcal{P}_3$	$\mathcal{P}_3$	$\mathcal{P}_3$	$\mathcal{P}_3$	$\mathcal{P}_3$	$\mathcal{P}_3$
$p_3(p)$	$\mathcal{P}_8$	$\mathcal{P}_8$	$\mathcal{P}_8$	$\mathcal{P}_8$	$\mathcal{P}_8$	$\mathcal{P}_8$	$\mathcal{P}_8$	$\mathcal{P}_8$	$\mathcal{P}_8$	$\mathcal{P}_8$	$\mathcal{P}_8$
$p_4(p)$	$\mathcal{P}_7$	$\mathcal{P}_7$	$\mathcal{P}_7$	$\mathcal{P}_7$	$\mathcal{P}_7$	$\mathcal{P}_7$	$\mathcal{P}_7$	$\mathcal{P}_7$	$\mathcal{P}_7$	$\mathcal{P}_7$	$\mathcal{P}_7$
$b_1(l)$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$
$b_1(r)$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$	$\mathcal{P}_1$
$b_2(l)$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$
$b_2(r)$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$	$\mathcal{P}_4$
$b_3(l)$	–	–	$\mathcal{P}_5$	–	–	–	$\mathcal{P}_8$	–	–	–	–
$b_3(r)$	–	–	$\mathcal{P}_6$	–	–	$\mathcal{P}_8$	–	–	–	–	–
$b_4(l)$	–	–	–	$\mathcal{P}_7$	–	–	–	–	–	–	$\mathcal{P}_7$
$b_4(r)$	–	–	–	$\mathcal{P}_7$	–	–	–	–	–	$\mathcal{P}_7$	–
$b_5(l)$	–	–	–	$\mathcal{P}_9$	–	–	–	–	–	–	–
$b_5(r)$	–	–	–	$\mathcal{P}_{10}$	–	–	–	–	–	–	–
		di			**co**			**nc**	**nd**		

2.2 Behavior Space

Starting from its initial state $\mathcal{H}_0$, HDES $\mathcal{H}$ may perform a sequence of component transitions within its *behavior space*, written $Bsp(\mathcal{H}, \mathcal{H}_0)$, which is a finite automaton

$$Bsp(\mathcal{H}, \mathcal{H}_0) = (\mathbf{S}, \mathbf{T}, S_0).$$

$\mathbf{S}$ is the set of states $(\mathcal{S}, \mathcal{E}, \mathcal{P})$, with $\mathcal{S} = (s_1, \ldots, s_n)$ being the tuple of states of components in $\mathcal{H}$. $\mathcal{E} = (e_1, \ldots, e_m)$ is the tuple of events at input terminals of components in $\mathcal{H}$ (ϵ indicates no event), and $\mathcal{P} = (P_1, \ldots, P_k)$ the tuple of pattern-space states. $S_0 = (\mathcal{H}_0, \mathcal{E}_0, \mathcal{P}_0)$ is the initial state, where $\mathcal{E}_0 = (\epsilon, \ldots, \epsilon)$ and $\mathcal{P}_0 = (P_{10}, \ldots, P_{k0})$ the tuple of the initial states of pattern spaces $Pts(X_1), \ldots, Pts(X_k)$, respectively. $\mathbf{T}$ is the transition function, where

$$(\mathcal{S}, \mathcal{E}, \mathcal{P}) \xrightarrow{T} (\mathcal{S}', \mathcal{E}', \mathcal{P}') \in \mathbf{T} \text{ if and only if:}$$

1. $T = s \xrightarrow{(e,I) \mid E_{\text{out}}} s'$ (where e is the input event and E_{out} the output events with relevant terminals) is a transition of a component C such that s equals one element of $\mathcal{S}$, I is an input terminal of C, and $\mathcal{E}(I) = e$;
2. $\mathcal{S}'$ differs from $\mathcal{S}$ only in s' replacing s;
3. If $C \in \mathcal{C}(X_i)$, $i \in [1..k]$, then $\mathcal{P}'$ differs from $\mathcal{P}$ only in the i-th element as follows:
$$\mathcal{P}'(P_i) = \begin{cases} \bar{P} & \text{if } \mathcal{P}(P_i) \xrightarrow{T} \bar{P} \in Pts(X_i) \\ P_{i0} & \text{otherwise;} \end{cases}$$
4. $\mathcal{E}'$ differs from $\mathcal{E}$ based on these conditions:

(*a*) $\mathcal{E}'(I) = \epsilon$ (event e is consumed);
(*b*) $\forall(o, O) \in E_{\text{out}}$, $\mathcal{E}(I') = \epsilon$, $\mathcal{E}'(I') = o$, where I' is the terminal entered by the link exiting O;
(*c*) If $\mathcal{P}'(P_i) \neq \mathcal{P}(P_i)$, $i \in [1\,..\,k]$, $\mathcal{P}'(P_i)$ is final in $Pts(X_i)$ and marked by complex event $\mathbf{p}$, then $\mathcal{E}(Cot) = \epsilon$, $\mathcal{E}'(Cot) = \mathbf{p}$.

As such, transitions in $Bsp(\mathcal{H}, \mathcal{H}_0)$ are marked by transitions of components in $\mathcal{H}$. The new state not only reflects the consumption of input event e of the component transition T and the generation of the output events in E_{out}: it also accounts for the possible occurrence of a complex event $\mathbf{p}$.

A string in the language of $Bsp(\mathcal{H}, \mathcal{H}_0)$ is a *history* of $\mathcal{H}$. The behavior space is defined for formal reasons only, as its actual materialization is impractical in real HDESs.

3 Problem Formulation

Diagnosing a HDES means finding the faults in its history. A history can be observed only in its observable transitions, as a sequence of *observation labels*, called the *trace* of the history, with each label being associated with an observable transition. The diagnosis process is complicated by two facts. First, several histories may generate the same trace. Second, because of noise and distribution of the channels conveying labels from the HDES, rather than a sequence of labels, the trace is perceived as a DAG, called *temporal observation*, where each node contains a set of observation labels and each arc represents partial (rather than total) temporal ordering between observation labels. Consequently, several *candidate traces* are observed, each one being made up by choosing a label in each node of the DAG without violating the temporal constraints imposed by arcs. Furthermore, since several (even infinite) histories may be consistent with the same trace, the diagnosis output is a set of *candidate diagnoses*, with each candidate corresponding to a subset of the possible histories. However, despite the possible infinite set of histories consistent with the temporal observation, the set of candidate diagnoses is always finite (being it upper-bounded by the powerset of component transitions).

A *diagnosis problem* for a HDES $\mathcal{H}$ is a quadruple

$$\wp(\mathcal{H}) = (\mathcal{H}_0, \mathcal{V}, \mathcal{O}, \mathcal{R}), \text{ where:}$$

- $\mathcal{H}_0$ is the initial state of $\mathcal{H}$;
- $\mathcal{V}$ is the *viewer* of $\mathcal{H}$, a set of pairs (T, ℓ), where T is a transition and ℓ an observation label, with $h_{[\mathcal{V}]}$ denoting the trace of history h based on $\mathcal{V}$;
- $\mathcal{O}$ is the *temporal observation* of $\mathcal{H}$, with $\|\mathcal{O}\|$ denoting the set of candidate traces;
- $\mathcal{R}$ is the *ruler* of $\mathcal{H}$, a set of pairs (T, f), where T is a transition and f a fault label, with $h_{[\mathcal{R}]}$ denoting the *diagnosis* of history h based on $\mathcal{R}$, defined as:

$$h_{[\mathcal{R}]} = \{f \mid T \in h, (T, f) \in \mathcal{R}\}.$$

If a transition T is involved in $\mathcal{V}$, then it is *observable*, otherwise it is *unobservable*. If T is involved in $\mathcal{R}$, then it is *faulty*, otherwise it is *normal*.

The *solution* Δ of $\wp(\mathcal{H})$ is the set of candidate diagnoses:

$$\Delta(\wp(\mathcal{H})) = \left\{\delta \mid h \in Bsp(\mathcal{H}, \mathcal{H}_0), h_{[\mathcal{V}]} \in \|\mathcal{O}\|, \delta = h_{[\mathcal{R}]}\right\}.$$

Each candidate diagnosis is the set of faulty transitions of a history that is consistent with the temporal observation.

For practical reasons, instead of processing the temporal observation $\mathcal{O}$, the *index space* of $\mathcal{O}$ is generated, namely $Isp(\mathcal{O})$. This is a deterministic automaton whose language equals $\|\mathcal{O}\|$ (the set of candidate traces).

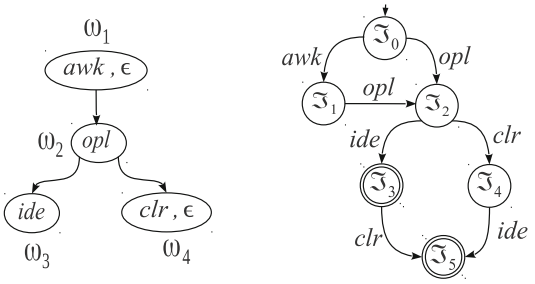


Fig. 3. Temporal observation $\mathcal{O}$ (left) and relevant index space $Isp(\mathcal{O})$ (right)

Example 3. With reference to Example 1, we define the diagnostic problem for the left-hand side protection-hardware as $\wp(\mathcal{W}) = (\mathcal{W}_0, \mathcal{V}, \mathcal{O}, \mathcal{R})$, where:

- In $\mathcal{W}_0$ both breakers are *closed*, protection device is *idle*, and protection hardware is *con* (see states in Fig. 2);
- $\mathcal{V} = \{(b_1(l), opl), (b_1(r), opr), (b_2(l), cll), (b_2(r), clr), (p_1, awk), (p_2, ide), (p_3, awk), (p_4, ide)\}$;
- $\mathcal{O}$ is the temporal observation displayed in Fig. 3 (left);
- $\mathcal{R} = \{(b_3(l), nol), (b_3(r), nor), (b_4(l), ncl), (b_4(r), ncr), (p_3, fop), (p_4, fcp), (w_1, fdw), (w_6, fcw), (w_7, fcw), (l_1, fil), (l_2, frl), (l_3, psl)\}$,
  with fault labels having the following meaning: $nol : l$ fails to open; $nor : r$ fails to open; $ncl : l$ fails to close; $ncr : r$ fails to close; $fop : p$ fails to trip breakers to close; $fcp : p$ fails to trip breakers to open; $fdw : \mathcal{W}$ fails to disconnect the line; $fcw : \mathcal{W}$ fails to connect the line; $fil : \mathcal{L}$ fails to be isolated; $frl : \mathcal{L}$ fails to be reconnected; $psl : \mathcal{L}$ is struck by a persistent short circuit.

Temporal observation $\mathcal{O}$ (Fig. 3, left) includes four nodes, with ω_1 and ω_4 containing two observation labels (ϵ is the empty label). Because of this uncertainty and partial temporal ordering, $\mathcal{O}$ embodies six candidate traces, which are the strings of the language of $Isp(\mathcal{O})$ displayed on the right of Fig. 3 (where $\Im_3$ and $\Im_5$ are final). $\diamond$

4 Diagnosis Computation

The definition of diagnosis-problem solution is not operational in nature: it refers to the behavior space, which is assumed not to be available in practice. The diagnosis engine is expected to be sound and complete in generating the solution of the problem, without the availability of the behavior space. To this end, it reconstructs only the subpart of the behavior space that is consistent with the temporal observation. In doing so, the reconstruction needs keeping four sorts of information: the state of components, the state of input terminals, the state of the matching of pattern events, and the state of the matching of the temporal observation. Specifically, the solution of a diagnostic problem $\wp(\mathcal{H}) = (\mathcal{H}_0, \mathcal{V}, \mathcal{O}, \mathcal{R})$ is computed in three steps:

- Generating the *index space* of temporal observation $\mathcal{O}$;
- Generating the subspace of $Bsp(\mathcal{H}, \mathcal{H}_0)$ that is consistent with temporal observation $\mathcal{O}$, based on viewer $\mathcal{V}$, called the *behavior* of $\wp(\mathcal{H})$, written $Bhv(\wp(\mathcal{H}))$;
- Decorating the states of behavior $Bhv(\wp(\mathcal{H}))$ by the associated set of candidate diagnoses.

The actual solution $\Delta(\wp(\mathcal{H}))$ is the union of the decorations associated with final states of $Bhv(\wp(\mathcal{H}))$ (see Theorem 1).

Formally, $Bhv(\wp(\mathcal{H}))$ is defined as follows. Let $\mathbf{S}$ be the domain of tuples $(s_1, \ldots, s_n)$ of states of components in $\mathcal{C}(\mathcal{H})$. Let $\mathbf{E}$ be the domain of tuples $(e_1, \ldots, e_m)$ of events at input terminals (other than *Ext*) of components in $\mathcal{C}(\mathcal{H})$. Let $\mathfrak{I}$ be the domain of states in $Isp(\mathcal{O})$. Let $\boldsymbol{\mathcal{P}}$ be the domain of tuples $(P_1, \ldots, P_k)$ of pattern-space states. The behavior of $\wp(\mathcal{H})$ is a deterministic automaton:

$$Bhv(\wp(\mathcal{H})) = (\mathbb{S}, \mathbb{T}, S_0, \mathbb{S}_\mathrm{f}), \text{ where}$$

- $\mathbb{S} \subseteq \mathbf{S} \times \mathbf{E} \times \boldsymbol{\mathcal{P}} \times \mathfrak{I}$ is the set of states;
- $S_0 = (\mathcal{H}_0, \mathcal{E}_0, \mathcal{P}_0, \mathfrak{I}_0)$ is the initial state, where $\mathcal{E}_0 = (\epsilon, \ldots, \epsilon)$, $\mathcal{P}_0 = (P_{10}, \ldots, P_{k0})$ the tuple of the initial states of pattern spaces $Pts(X_1), \ldots, Pts(X_k)$, respectively, and $\mathfrak{I}_0$ the initial state of $Isp(\mathcal{O})$;
- $\mathbb{S}_\mathrm{f} = \{(\mathcal{S}, \mathcal{E}, \mathcal{P}, \mathfrak{I}) \mid \mathcal{E} = (\epsilon, \ldots, \epsilon), \mathfrak{I} \text{ is final } \}$ is the set of final states;
- $\mathbb{T}$ is the transition function, where

$$(\mathcal{S}, \mathcal{E}, \mathcal{P}, \mathfrak{I}) \xrightarrow{T} (\mathcal{S}', \mathcal{E}', \mathcal{P}', \mathfrak{I}') \in \mathbb{T} \text{ if and only if:}$$

1. Conditions 1–4 on $\mathcal{S}'$, $\mathcal{E}'$, and $\mathcal{P}'$, in the specification of the transition function of $Bsp(\mathcal{H}, \mathcal{H}_0)$, hold;
2. $\mathfrak{I}' = \begin{cases} \bar{\mathfrak{I}} & \text{if } (T, o) \in \mathcal{V}, \mathfrak{I} \xrightarrow{o} \bar{\mathfrak{I}} \in Isp(\mathcal{O}) \\ \mathfrak{I} & \text{otherwise.} \end{cases}$

The actual algorithm that builds $Bhv(\wp(\mathcal{H}))$ starts from the initial state S_0 and generates all possible transitions based on the conditions above. Eventually, it removes all spurious states and transitions that are not in a path from the initial state to a final state.

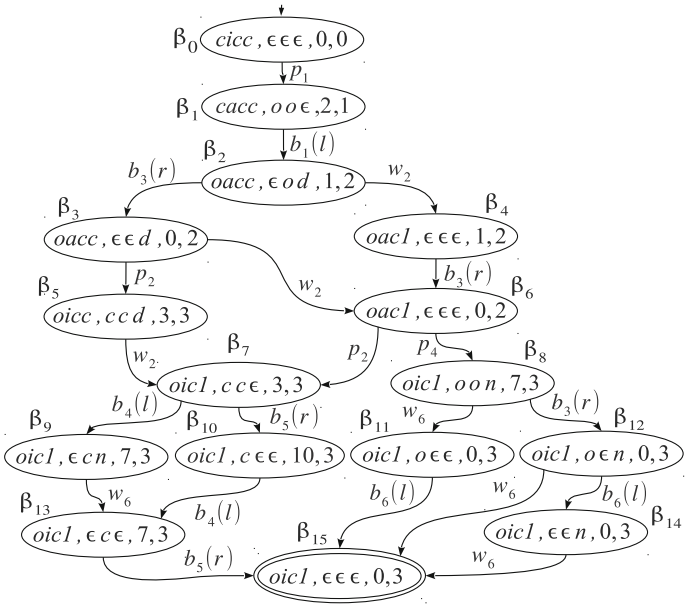


Fig. 4. Behavior $Bhv(\wp(\mathcal{W}))$

Example 4. With reference to $\wp(\mathcal{W})$ defined in Example 3, shown in Fig. 4 is $Bhv(\wp(\mathcal{W}))$, including states $\beta_0, \ldots, \beta_{15}$, with β_{15} final. Each state $(\mathcal{S}, \mathcal{E}, \mathcal{P}, \Im)$ is such that $\mathcal{S}$ is the quadruple of states for l, p, r, and $\mathcal{W}$, where *closed*, *open*, *idle*, *awaken*, *con*, and *dis1* are written c, o, i, a, c, and *1*, respectively, $\mathcal{E}$ is the triple of events at input terminals of l, r, and $\mathcal{W}$, respectively, where *op*, *cl*, **di**, and **nc** are written o, c, d, and n, respectively, while $\mathcal{P}$ and $\Im$ are the indices of states in $Pts(w)$ and $Isp(\mathcal{O})$, respectively. For instance, $\beta_2 = (oacc, \epsilon od, 1, 2)$ stands for $\mathcal{S} = (open, awaken, closed, con)$, $\mathcal{E} = (\epsilon, op, \mathbf{di})$, $\mathcal{P} = \mathcal{P}_1$, and $\Im = \Im_2$. ◇

Once generated the behavior, each state S of $Bhv(\wp(\mathcal{H}))$ is decorated by a set of candidate diagnoses $\Delta(S)$ based on the following two inductive rules:

(1) For the initial state: $\Delta(S_0) = \{\emptyset\}$.
(2) For each transition $S \xrightarrow{T} S'$ in $Bhv(\wp(\mathcal{H}))$:
   - If T is normal then $\delta \in \Delta(S) \Rightarrow \delta \in \Delta(S')$;
   - If $(T, f) \in \mathcal{R}$ then $\delta \in \Delta(S) \Rightarrow (\delta \cup \{f\}) \in \Delta(S')$.

The algorithm that decorates $Bhv(\wp(\mathcal{H}))$ starts by applying the first rule, marking the initial state with the singleton of the empty diagnosis. Then, based on the decoration of the initial state, it continuously applies the second rule for each transition exiting a state S whose decoration has changed. If $(T, f) \in \mathcal{R}$ then T is faulty, with f being the relevant fault. If so, each candidate diagnosis in $\Delta(S)$ extended by fault f is also a candidate diagnosis in $\Delta(S')$. Instead, if T is normal, all candidate diagnoses in $\Delta(S)$ are candidate diagnoses in $\Delta(S')$ too. As such, $\Delta(S)$ is constructed as the set of diagnoses

relevant to histories ending at state S. The algorithm terminates when the application of the second rule does not cause any change in any decoration.

Table 4. Decoration of $Bhv(\wp(\mathcal{W}))$

States	*Decoration*
$\beta_0, \beta_1, \beta_2, \beta_4$	$\{\emptyset\}$
$\beta_3, \beta_5, \beta_6, \beta_7, \beta_{10}$	$\{\{nor\}\}$
β_9	$\{\{nor, ncl\}\}$
β_{13}	$\{\{nor, ncl, fcw\}\}$
$\beta_8, \beta_{12}, \beta_{14}$	$\{\{nor, fcp\}\}$
β_{11}	$\{\{nor, fcp, fcw\}\}$
β_{15}	$\{\{nor, ncl, fcw\}, \{nor, fcp, fcw\}\}$

Example 5. Based on ruler $\mathcal{R}$ (Example 3), the behavior in Fig. 4 will be decorated as specified in Table 4. Therefore, two candidate diagnoses are associated with final state β_{15}, namely $\delta_1 = \{nor, ncl, fcw\}$, and $\delta_2 = \{nor, fcp, fcw\}$, corresponding to these two scenarios:

δ_1 : Breaker r fails to open, breaker l fails to close, and protection hardware $\mathcal{W}$ fails to connect;
δ_2 : Breaker r fails to open, protection device trips breakers to open rather than to close, and $\mathcal{W}$ fails to connect.

Based on Theorem 1, $\{\delta_1, \delta_2\}$ is the solution of $\wp(\mathcal{W})$. Albeit we have two candidates, since $\delta_1 \cap \delta_2 = \{nor, fcw\}$, certainly r failed to open and $\mathcal{W}$ failed to connect. $\diamond$

Theorem 1. *Let $\wp(\mathcal{H}) = (\mathcal{H}_0, \mathcal{V}, \mathcal{O}, \mathcal{R})$, and $\Delta(\mathcal{B}^v)$ denote the union of the sets of diagnoses decorating the final states of $Bhv(\wp(\mathcal{H}))$. Then, $\Delta(\wp(\mathcal{H})) = \Delta(\mathcal{B}^v)$.*

Proof. (*Sketch*) The proof is grounded on Lemmas 1–5, where $\mathcal{B}^s$ and $\mathcal{B}^v$ denote $Bsp(\mathcal{H}, \mathcal{H}_0)$ and $Bhv(\wp(\mathcal{H}))$, respectively.

Lemma 1. *If history $h \in \mathcal{B}^v$ then $h \in \mathcal{B}^s$.*

This derives from the fact $\mathcal{B}^v$ differs from $\mathcal{B}^s$ in the additional field $\Im$, which is irrelevant for conditions on $\mathcal{S}'$, $\mathcal{E}'$, and $\mathcal{P}'$. By induction on h, starting from the initial state, each new transition applicable in $\mathcal{B}^v$ is applicable in $\mathcal{B}^s$ too.

Lemma 2. *If history $h \in \mathcal{B}^v$ then $h_{[\mathcal{V}]} \in \|\mathcal{O}\|$.*

Recall that $h_{[\mathcal{V}]}$ is the sequence of observable labels associated with visible transitions in viewer $\mathcal{V}$. Based on the definition of $\mathcal{B}^v$, $h_{[\mathcal{V}]}$ belongs to the language of $Isp(\mathcal{O})$, which equals $\|\mathcal{O}\|$. Thus, $h_{[\mathcal{V}]} \in \|\mathcal{O}\|$.

Lemma 3. *If history* $h \in \mathcal{B}^s$ *and* $h_{[\mathcal{V}]} \in \|\mathcal{O}\|$ *then* $h \in \mathcal{B}^v$.

By induction on h, starting from the initial state, each new transition T applicable in $\mathcal{B}^s$ is applicable in $\mathcal{B}^v$ too. In fact, if T is invisible, no further condition is required. If T is visible, based on the assumption $h_{[\mathcal{V}]} \in \|\mathcal{O}\|$ and that the language of $Isp(\mathcal{O})$ equals $\|\mathcal{O}\|$, the label associated with T in viewer $\mathcal{V}$ matches a transition in $Isp(\mathcal{O})$.

Lemma 4. *If history* $h \in \mathcal{B}^v$ *ends at final state* S_f *then* $\Delta(S_\mathrm{f})$ *includes a candidate diagnosis* $h_{[\mathcal{R}]}$.

Based on the two rules for decoration of $\mathcal{B}^v$, by induction on h, starting from the initial state (h empty) and the empty diagnosis δ, the addition of a new transition T in h extends δ (within the decoration of the new state) by either nothing (T normal) or a fault label (T faulty). Upon the last transition of h, δ includes all fault labels associated with faulty transitions in ruler $\mathcal{R}$, in other words, $\delta = h_{[\mathcal{R}]}$.

Lemma 5. *If* S_f *is a final state in* $\mathcal{B}^v$ *and* $\delta \in \Delta(S_\mathrm{f})$ *then there exists a history* $h \in \mathcal{B}^v$ *ending at* S_f *such that* $h_{[\mathcal{R}]} = \delta$.

Based on the decoration rules for $\mathcal{B}^v$, δ is incrementally generated starting from the empty diagnosis initially associated with S_0, by inserting each faulty label associated with each faulty transition encountered in a path from S_0 to S_f. This path is a history provided that it is finite. In fact, cycles in $\mathcal{B}^v$ allow for an infinite number of applications of the second decoration rule. However, since δ is a set, once a cycle has been covered, all associated fault labels are inserted in δ. Successive iterations of the cycle do not extend δ because of duplicate removals. Thus, δ can always be generated by a finite history h, in other words, $\delta = h_{[\mathcal{R}]}$.

To prove Theorem 1, we show $\delta \in \Delta(\mathcal{B}^v) \Leftrightarrow \delta \in \Delta(\wp(\mathcal{H}))$. On the one hand, if $\delta \in \Delta(\mathcal{B}^v)$ then, based on Lemmas 1, 2, and 5, there exists a history $h \in \mathcal{B}^s$ such that $h_{[\mathcal{V}]} \in \|\mathcal{O}\|$ and $h_{[\mathcal{R}]} = \delta$, that is, $\delta \in \Delta(\wp(\mathcal{H}))$. On the other, if $\delta \in \Delta(\wp(\mathcal{H}))$ then, based on Lemmas 3 and 4, there exists a history $h \in \mathcal{B}^v$ ending at final state S_f such that $\delta = h_{[\mathcal{R}]}$ and $\delta \in \Delta(S_\mathrm{f})$, that is, $\delta \in \Delta(\mathcal{B}^v)$. □

5 Discussion

HDESs are a means of modeling complex DESs, where behavior is stratified and events can be generated by patterns of transitions. In spite of being influenced by other components, each internal node X of the hierarchy is a complex component living its own life. This means that X has its own behavioral model, which does not coincide with the composition of the behavioral models of its components. This results in a hierarchical system (HDES) made up of several cohabiting subsystems accommodated at different abstraction levels.

The diagnosis technique defined for HDESs is model-based in nature: diagnosis is output based on the model of the system and the temporal observation. Only the portion of the behavior space consistent with the observation is reconstructed and eventually

decorated by candidate diagnoses. Not only does separation of concerns apply to the modeling, it also applies to the diagnosis task. Since each (complex) component is provided with its own behavioral model, diagnosis is context-sensitive [17].

Moreover, depending on the degree of constraints on computational resources and time response of the diagnosis engine, model-based reasoning can be scaled to a convenient level of abstraction. This means restricting the HDES $\mathcal{H}$ to a portion $\mathcal{H}'$ (for instance, a complex component along with its children) and projecting observation $\mathcal{O}$ on $\mathcal{O}'$, resulting from the removal of irrelevant labels, nodes, and arcs. This way, within the context of $\mathcal{H}'$, the diagnosis output is complete, even if not sound (due to the removal of the behavioral constraints imposed by $\mathcal{H} - \mathcal{H}'$ and the observation constraint imposed by $\mathcal{O} - \mathcal{O}'$). To refine diagnosis, both $\mathcal{H}'$ and $\mathcal{O}'$ may be then enlarged to a suitable extent. Ideally, several restrictions $\mathcal{H}'_1, \ldots, \mathcal{H}'_n$ of $\mathcal{H}$ can be considered and diagnosed in parallel, with eventual combination of the reconstructed behaviors.

An additional benefit in modeling complex systems as HDESs is support to *integration*. If we have several DESs, each one devoted to a specific task, and we want to integrate them into a new DES in order to have control on all DESs at a higher abstraction level, then the new DES can be conveniently modeled as an HDES, which is sensitive to specific behavioral patterns of the DESs (complex events). This way, monitoring and diagnosing the HDES is far more natural than interpreting (possibly overwhelming) streams of low-level events generated by DESs, such as alarms detected in a control room. The key point is that the DESs are integrated in a new system not just by connecting them to one another by means of new links between components belonging to different DESs. Instead, besides this *physical integration*, *logical integration* is achieved too, so that the new HDES is not just the union of the DESs: it is an higher-level system with its proper behavior and, as such, living its own life.

6 Related Work

This paper substantially extends the idea of context-sensitive diagnosis [17] in three directions. First, in [17] pattern stratification is only apparent, as, after macro-substitution, the regular expression is invariably defined on (basic) component transitions. In this paper, pattern stratification is real, since regular expressions are defined on the transitions of possibly complex components. Second, in [17] faults are associated with pattern matching. In this paper, faults are associated with transitions of components: pattern matching generates pattern events and, by union, complex events, to which complex components are sensitive. More importantly, in [17] context-sensitivity is defined on active systems, while this paper deals with HDESs, which provide behavioral stratification: separation of concerns holds not only for diagnosis but also for behavior.

This paper also differs from [13], where the notion of supervision pattern is introduced, mainly because neither a hierarchical structure for the system is conceived nor behavioral stratification is applicable.

HDESs are not HFSMs (*Hierarchical Finite State Machines*). The notion of an HFSM was inspired by statecharts [10, 11], a visual formalism for complex systems. The most important feature of an HFSM is hierarchical state-nesting: if a system is in a nested state (substate), it is also in all its surrounding states (superstates). Moreover,

transitions are defined at each level of the hierarchy. This resembles the idea of class inheritance in object-orientation and, in fact, it provides relevant advantages, including factorization and reuse. A simplified version of statechart, namely HFSM, was considered for solving a class of control problems in [4]. Recently, diagnosis of HFSMs has been considered in [12, 19]. However, no patterns are involved, events are simple, and diagnosis is context-free.

7 Conclusion

HDESs are a means to formalize complex DESs with behavior stratification. This allows for the modeling of a hierarchy wherein different, yet integrated, subsystems coexist, each one living its own life. Benefits include context-sensitivity and scalability of diagnosis, as well as support to logical system-integration. It is our belief that monitoring and diagnosing complex systems, such as a nuclear plant or a large power network, requires some sort of abstraction and separation of concerns. The ideas presented in this paper may be a step in the right direction.

Acknowledgment. This work was supported in part by NSFC under Grant No. 61003101, and Zhejiang Provincial Natural Science Foundation under Grant No. Y1100191.

References

1. Report of the Enquiry Committee on Grif Disturbance in Northern Region on 30th July 2012 and in Northern, Eastern & North-Eastern Region on 31st July 2012 (2012), http://www.powermin.nic.in/pdf/GRID_ENQ_REP_16_8_12.pdf
2. Baroni, P., Lamperti, G., Pogliano, P., Zanella, M.: Diagnosis of large active systems. Artificial Intelligence 110(1), 135–183 (1999)
3. Brand, D., Zafiropulo, P.: On communicating finite-state machines. Journal of ACM 30(2), 323–342 (1983)
4. Brave, H., Heymann, M.: Control of discrete event sysytems modeled as hierarchical state machines. IEEE Transactions on Automatic Control 38(12), 1803–1819 (1993)
5. Cassandras, C., Lafortune, S.: Introduction to Discrete Event Systems. The Kluwer International Series in Discrete Event Dynamic Systems, vol. 11. Kluwer Academic Publishers, Boston (1999)
6. Debouk, R., Lafortune, S., Teneketzis, D.: Coordinated decentralized protocols for failure diagnosis of discrete-event systems. Journal of Discrete Event Dynamic Systems: Theory and Applications 10, 33–86 (2000)
7. Debouk, R., Lafortune, S., Teneketzis, D.: On the effect of communication delays in failure diagnosis of decentralized discrete event systems. Journal of Discrete Event Dynamic Systems: Theory and Applications 13, 263–289 (2003)
8. Grastien, A., Cordier, M., Largouët, C.: Extending decentralized discrete-event modelling to diagnose reconfigurable systems. In: Carcassonne, F. (ed.) Fifteenth International Workshop on Principles of Diagnosis –DX 2004, pp. 75–80 (2004)

9. Grastien, A., Cordier, M., Largouët, C.: Incremental diagnosis of discrete-event systems. In: Sixteenth International Workshop on Principles of Diagnosis DX 2005, Monterey, CA, pp. 119–124 (2005)
10. Harel, D.: Statecharts: a visual formalism for complex systems. Science of Computer Programming 8, 231–274 (1987)
11. Harel, D., Lachover, H., Naamad, A., Pnueli, A., Politi, M., Sherman, R., Shtull-Trauring, A., Trakhtenbrot, M.: STATEMATE: a working environment for the development of complex reactive systems. IEEE Transactions on Software Engineering 16(4), 403–414 (1990)
12. Idghamishi, A., Zad, S.: Fault diagnosis in hierarchical discrete-event systems. In: 43rd IEEE Conference on Decision and Control, pp. 63–68. Paradise Island, BSH (2004)
13. Jéron, T., Marchand, H., Pinchinat, S., Cordier, M.: Supervision patterns in discrete event systems diagnosis. In: Peñaranda de Duero, E. (ed.) Seventeenth International Workshop on Principles of Diagnosis DX 2006, pp. 117–124 (2006)
14. Kwong, R., Yonge-Mallo, D.: Fault diagnosis in discrete-event systems: incomplete models and learning. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 41(1), 118–130 (2011)
15. Lamperti, G., Zanella, M.: Diagnosis of discrete-event systems from uncertain temporal observations. Artificial Intelligence 137(1–2), 91–163 (2002)
16. Lamperti, G., Zanella, M.: Diagnosis of Active Systems – Principles and Techniques. The Kluwer International Series in Engineering and Computer Science, vol. 741. Kluwer Academic Publishers, Dordrecht (2003)
17. Lamperti, G., Zanella, M.: Context-sensitive diagnosis of discrete-event systems. In: Walsh, T. (ed.) Twenty-Second International Joint Conference on Artificial Intelligence IJCAI 2011, pp. 969–975. AAAI Press, Barcelona (2011)
18. Lamperti, G., Zanella, M.: Monitoring of active systems with stratified uncertain observations. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 41(2), 356–369 (2011)
19. Paoli, A., Lafortune, S.: Diagnosability analysis of a class of hierarchical state machines. Journal of Discrete Event Dynamic Systems: Theory and Applications 18(3), 385–413 (2008)
20. Pencolé, Y.: Decentralized diagnoser approach: application to telecommunication networks. In: Eleventh International Workshop on Principles of Diagnosis DX 2000, Morelia, MX, pp. 185–192 (2000)
21. Pencolé, Y., Cordier, M.: A formal framework for the decentralized diagnosis of large scale discrete event systems and its application to telecommunication networks. Artificial Intelligence 164, 121–170 (2005)
22. Pencolé, Y., Cordier, M., Rozé, L.: Incremental decentralized diagnosis approach for the supervision of a telecommunication network. In: San Sicario, I. (ed.) Twelfth International Workshop on Principles of Diagnosis DX 2001, pp. 151–158 (2001)
23. Qiu, W., Kumar, R.: Decentralized failure diagnosis of discrete event systems. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 36(2), 384–395 (2006)
24. Yamaguchi, M.: Fukushima Nuclear Disaster Report: Plant Operators Tokyo Electric And Government Still Stumbling (2012), http://www.hungtonpost.com/2012/07/23/fukushima-dai-ichi-nuclear-plant-operators_n_1694476.html
25. Zhao, X., Ouyang, D.: Model-based diagnosis of discrete event systems with an incomplete system model. In: Eigtheenth European Conference on Artificial Intelligence ECAI 2008, pp. 189–193. IOS Press, Amsterdam (2008)
26. Zhao, X., Ouyang, D., Zhang, L., Wang, X., Mo, Y.: Reasoning on partially-ordered observations in online diagnosis of DESs. AI Communications 25(4), 285–294 (2012)

Combining Goal-Oriented and Problem-Oriented Requirements Engineering Methods*

Kristian Beckers[1], Stephan Faßbender[1], Maritta Heisel[1], and Federica Paci[2]

[1] Paluno - The Ruhr Institute for Software Technology – University of Duisburg-Essen
{firstname.lastname}@paluno.uni-due.de
[2] Department and Information Engineering and Computer Science, University Trento
{firstname.lastname}@unitn.it

Abstract. Several requirements engineering methods exist that differ in their abstraction level and in their view on the system-to-be. Two fundamentally different classes of requirements engineering methods are goal- and problem-based methods. Goal-based methods analyze the goals of stakeholders towards the system-to-be. Problem-based methods focus on decomposing the development problem into simple sub-problems. Goal-based methods use a higher abstraction level that consider only the parts of a system that are relevant for a goal and provide the means to analyze and solve goal conflicts. Problem-based methods use a lower abstraction level that describes the entire system-to-be. A combination of these methods enables a seamless software development, which considers stakeholders' goals and a comprehensive view on the system-to-be at the requirements level. We propose a requirements engineering method that combines the goal-based method SI* and the problem-based method Problem Frames. We propose to analyze the issues between different goals of stakeholders first using the SI* method. Our method provides the means to use the resulting SI* models as input for the problem frame method. These Problem Frame models can be refined into architectures using existing research. Thus, we provide a combined requirements engineering method that considers all stakeholder views and provides a detailed system specification. We illustrate our method using an E-Health example.

Keywords: requirements engineering, SI*, Problem Frames.

1 Introduction

Eliciting and analyzing requirements of a system is important, because it is hard to built the right software if you do not know what right is. Several methods exist that support the eliciting and analyzing requirements. Fabian et al. [8] present a survey of security requirements engineering (SRE) methods and one of their finding is that the integration of different SRE is worthwhile. The reason is that these all produce different artifacts and support different views and abstraction levels. Hence, a combination of these results can improve the quality of the requirements.

* This research was partially supported by the EU project Network of Excellence on Engineering Secure Future Internet Software Services and Systems (NESSoS, ICT-2009.1.4 Trustworthy ICT, Grant No. 256980).

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 178–194, 2013.

We propose to combine goal-based and problem-based requirements engineering methods. We use the SI* modeling language [18] as an example for a goal-based method and problem frames [12] as an example for a problem-based requirements engineering method.

The goal-based methods consider the views towards the system-to-be via its stakeholders' goals. The software to be developed is considered only as needed to achieve the goals. For example, on the one hand, a hospitals wants to provide treatment for a patient and requires monitoring data of the patient to administer the treatment. In this case only the monitoring data is modeled. This, however, is sufficient to analyze possible goal conflicts. On the other hand, a researcher wants to use this monitoring data to conduct a medical study. This can be in violation against the privacy goals of the patient. Hence, goal-based methods provide the means to detect and resolve these conflicts. However, the gap to a design description of the overall system is significant, because the system-to-be is only modelled in some artifacts that have a relation to a goal.

The problem-based methods follow a fundamentally different approach. These methods focus on the problem that shall be solved by the system-to-be. Problem-based methods model the environment and the system-to-be, called *machine*, in it. They describe requirements as an effect the system-to-be has on the environment. Refinements allow one to split the problem into sub-problems. Problem-based methods provide a description of the machine and the interaction with its environment. The sub-problems can be further refined into a design description and a system architecture, e.g., via the *ADIT* method [6,9]. Hence, the abstraction level of problem-based methods is lower than the abstraction level of goal-based methods.

Problem-based methods do not consider the views of all stakeholders and do not consider goals. Thus, they cannot detect or resolve goal-based conflicts. We propose to combine SI* with problem frames in order to create a method that considers all stakeholder views and can resolve goal-based conflicts, while also creating a description of the machine to be built and a structured way to create an architecture.

The rest of the paper is organized as follows. In the next section we discuss related work. In Section 3, we introduce the problem frame method and the SI* modeling language. Section 4 presents our method for seamless software development from goals to software architectures. In Section 5, we apply our method to the E-Health example. Section 6 discusses our results and Section 7 concludes the paper.

2 Related Work

Massacci et al. [17] have investigated the relations between Secure Tropos, Problem Frames and general security concepts e.g. assets. They have proposed a unified ontology to reach a shared understanding of the domain of security requirements, and also take advantage of multiple techniques to model and analyze security requirements. The ontology amalgamates the security ontologies of SI* [18] and Problem Frames [12], and accounts for rather nebulous security concepts, such as those of vulnerability and threat. The method differs from our own, because the authors do not investigate how to combine these methods but just explore the relationship between SI* and Problem Frames concepts.

Fabian et al. [8] propose a conceptual framework for security requirements engineering. The authors also define a common terminology for the framework. The purpose of the framework is to compare security requirements engineering methods. For example, goals are one concept in the framework, and the authors investigate if methods support this concept or not and if the concept is used under a different name. Our method uses the idea of this work that requirements are a refinement of goals and that the SI* method uses a higher abstraction level than Problem Frames. However, the authors do not show a method that actually combines methods.

Liu and Jin [15] propose relations between SI* and Problem Frames. The authors propose to introduce a domain actor and domain constraints into SI* in order to link SI* and problem frame models. This method differs from our own, because we propose to use the methods in sequence, first SI* than Problem Frames. The authors use both methods at the same time.

Supakkul and Chung [19] introduce the concept of a soft-goal from SI* into the Problem Frame method. This method provides support for modeling stakeholder goals in Problem Diagrams. This method differs from our own, because we try to use both methods and not to integrate one into the other. The advantage is that a development is analysed separately on different abstraction layers.

Classen et al. [5] propose to integrate feature diagrams and the Problem Frame method. The authors propose to use these methods in sequence. The Problem Frame method is used first for requirements elicitation and analysis. Feature diagrams contain the solutions for the elicited requirements. The authors also do not change the existing methods and integrate problem frames with another existing method. This method can complement our own.

Letier and van Lamsweerde [14] propose a method to construct software specification from high-level goals. The method uses a formal patterns that guide the operationalization. The method differs from our own, because it results in a specification of pieces of the software. Our work results in the specification of the entire architecture.

3 Background

We introduce the notations we combine in this work in this section. We explain the goal-based notation SI* in Sect. 3.1 and the problem-based notation Problem Frames in Sect. 3.2.

3.1 SI*

The SI* modeling language has been proposed to capture security and functional requirements of socio-technical systems. SI* is founded on the concepts of *agent*, *role*, *goal*, *task*, *resource*. An agent is an active entity with concrete manifestations and is used to model humans as well as software agents and organizations. A role is the abstract characterization of the behavior of an active entity within some context. They are graphically represented as circles. Assignments of agents to roles are described by the *play* relation.[1] A goal is a state of affairs whose realization is desired by some actor

[1] For the sake of simplicity, in the remainder of the paper we use the term *actor* to indicate agents and roles when it is not necessary to distinguish them.

(objective), can be realized by some (possibly different) actor (capability), or should be authorized by some (possibly different) actor (entitlement). Entitlements, capabilities and objectives of actors are modeled through relations between an actor and a goal: *own* indicates that an actor has full authority concerning ac- cess and disposition over his entitlement; *provide* indicates that an actor has the capabilities to achieve the goal; and *request* indicates that an actor intends to achieve the goal. A task specifies the procedure used to achieve goals. A resource represents a physical or an informational entity without intentionality. A resource can be consumed or produced by a task. In the graphical representation, goals,tasks and resources are respectively represented as ovals, hexagons, and rectangles. Own, provide, and request are represented with edges between an actor and a goal labeled by **O**, **P**, and **R**, respectively. Goals and tasks of the same actor or of different actors are often related to one another in many ways. AND/OR decomposition combines AND and OR refinements of a root goal into subgoals. However, neither such goals might be under the control of the actor nor the actor may have the capabilities to achieve them. *Contribution* relations are used when the relation between goals is not the consequence of a deliberative planning but rather results from side-effects. The impact can be positive or negative and is graphically represented as edges labeled with + and -, respectively. Finally, tasks are linked to the goals that they intend to achieve using *means-end* relations. The relations between actors within the system are captured by the notions of *delegation* and *trust*. Assignment of responsibilities among actors can be made by *execution dependency* (when an actor depends on another actor for the achievement of a goal) or *permission delegation* (when an actor authorizes another actor to achieve the goal). Usually, an actor prefers to appoint actors that are expected to achieve assigned duties and not misuse granted permissions. SI* adopts the notions of *trust of execution* and *trust of permission* to model such expectations. In the graphical representation, permission delegations are represented with edges labeled by **Dp** and execution dependencies with edges labeled by **De**. Finally, trust of permission relations are represented with edges labeled by **Tp** and trust of execution relations with edges labeled by **Te**.

3.2 Problem Frames

Problem frames are a means to describe software development problems. They were proposed by Jackson [12], who describes them as follows: *"A problem frame is a kind of pattern. It defines an intuitively identifiable problem class in terms of its context and the characteristics of its domains, interfaces and requirement."*. It is described by a *frame diagram*, which consists of domains, interfaces between them, and a requirement. We describe problem frames using class diagrams extended by stereotypes as proposed by Hatebur and Heisel [11]. All elements of a problem frame diagram act as placeholders, which must be instantiated to represent concrete problems. Doing so, one obtains a problem description that belongs to a specific class of problems.

Figure 1 shows an example of a problem frame. The class with the stereotype machine represents the thing to be developed (e.g., the software). The classes with some domain stereotypes, e.g., causalDomain or biddableDomain represent *problem domains* that already exist in the application environment. Jackson distinguishes the domain types *causal domains* that comply with some physical laws, *lexical domains* that are

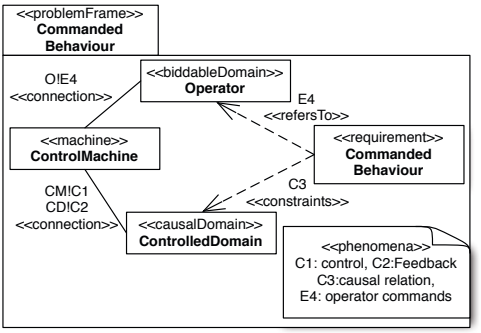


Fig. 1. *Commanded Behaviour* problem frame

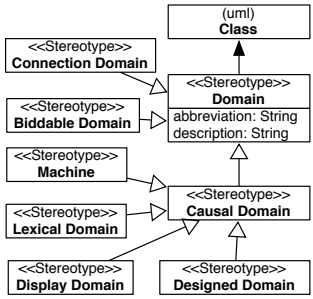


Fig. 2. Inheritance structure of different domain types

data representations, and *biddable domains* that are usually people. We use the formal meta model [11] shown in Fig. 2 to annotate domains with their corresponding stereotype.

Domains are connected by interfaces consisting of shared phenomena. Shared phenomena may be events, operation calls, messages, and the like. They are observable by at least two domains, but controlled by only one domain, as indicated by an exclamation mark. For example, in Fig. 1 the notation *O!E4* means that the phenomena in the set *E4* are controlled by the domain ***Operator***. These interfaces are represented as associations, and the name of the associations contain the phenomena and the domains controlling the phenomena.

In Fig. 1, the ***ControlledDomain*** domain is constrained and the ***Operator*** is referred, because the ***ControlMachine*** has the role to change the ***ControlledDomain*** on behalf of the ***Operator***'s commands for achieving the required ***Commanded Behaviour***. These relationships are modeled using dependencies that are annotated with the corresponding stereotypes.

Problem frames support developers in analyzing problems to be solved. They show what domains have to be considered, and what knowledge must be described and reasoned about when analyzing the problem in depth. Other basic problem frames besides the commanded behavior frame shown in Fig. 1 are *required behaviour*, *simple workpieces*, *information display*, and *transformation* [12].

Software development with problem frames proceeds as follows: first, the environment in which the machine will operate is represented by a *context diagram*. Like a frame diagram, a context diagram consists of domains and interfaces. However, a context diagram contains no requirements. Then, the problem is decomposed into subproblems. If ever possible, the decomposition is done in such a way that the subproblems fit to given problem frames. To fit a sub-problem to a problem frame, one must instantiate its frame diagram, i.e., provide instances for its domains, phenomena, and interfaces. The instantiated frame diagram is called a *problem diagram*.

Since the requirements refer to the *environment* in which the machine must operate, the next step consists in deriving a *specification* for the machine (see [13] for details). The specification describes the machine and is the starting point for its construction.

Problem frames are an appropriate means to analyze not only functional, but also dependability and other quality requirements [10,1].

The UML4PF framework provides tool support for this method. A more detailed description can be found in [7].

4 A Method for Goal- and Problem-Based Software Engineering

In this section we present our method for combining goal- and problem-based software engineering in Sect. 4.1. The method uses a mapping from the SI* notation to the Problem Frame notation, which is shown in Sect. 4.2. Section 4.3 presents some consistency checks for our method.

4.1 Method

We propose the following method that integrates the SI* and the Problem Frames method, depicted in Fig. 3.

1. **SI* model instantiation.** The aim of this step is to draw an SI* model that captures the system's stakeholders goals. This means identifying the actors and their goals of the system-to-be. The goals are analysed next and all resources or tasks that are a means to to fulfil them are included in the model, as well. It is also possible that one stakeholder requires the resource of another to fulfil his/her goal. In this case, we need to add trust relationships, so that one stakeholder is entrusted with the usage of the resource of another (see Sect. 3).
2. **Goal conflict analysis.** The second step of our method conducts a conflict analysis on the SI* models and resolves these conflicts. We analyse the impacts one goal has on another. The first question, has a goal any impact on another. If this is the case, we have to analyse if one goal contributes to fulfil another or if it is an obstacle for fulfilling it. The resulting relation is included into the SI* model as described in Sect. 3. It results in an evolved SI* models, which contain the information how goals relate to one another.

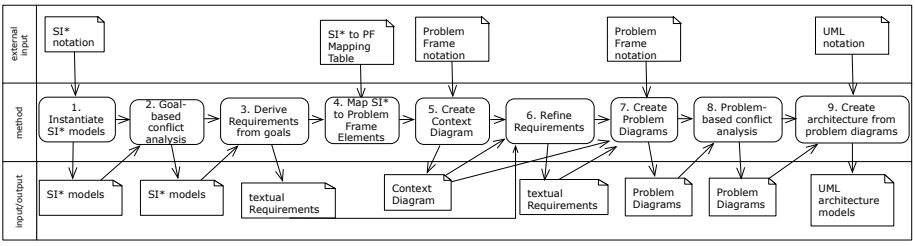


Fig. 3. A method for the integration of SI* and Problem Frames

3. **Derive requirements from goals.** The third step demands to write down requirements, which are based upon the goals in the SI* models. We analyse each goal of every stakeholder and formulate a text that consider all the information in the SI* models relevant for that particular goal. For example, we should name the stakeholder the goal belongs to in the text, the resources and tasks required to fulfil it and further relations to other stakeholders, e.g., trust relations.
4. **Mapping SI* to Problem Frames.** This step consists of the mapping of the elements in the SI* models to elements of the problem frame notations. We use Tab. 1 to guide the mapping. The table will be explained in the following subsection. The table states what elements of problems to consider, when mapping an SI* element. This requires in several cases decision by a human expert, because the table does not present a one to one mapping, but offers choices of problem frame notation for most of the SI* elements. It is critical for the requirements engineer to understand the meaning behind every instantiated SI* element in order to execute the mapping. This shall also stimulate a discussion between the customer and the requirements that helps understanding the system-to-be better and validates the instantiated SI*. This validation is based on the knowledge of the customer and is not to be confused with the validation executed by the requirements engineer alone, which is described in Sect. 4.3.
5. **Create context diagram.** We use the mapping of the previous step to create our fist diagram using the problem frame notation. This diagram is our context diagram (see Sect. 3), which describes the context of the development problem. While the SI* models model the actors and their goals, and resources were shown only if they can fulfil a goal. Problem frames is based on the machine to be build. Hence, we have to create the machine domain first and check if it already exist in the set of domains resulting from the mapping in the previous step. If not we have to add it to the context diagram. Next, we add the domains, which whom the machine domain has a relation and exchanges phenomena. We check the set of domains resulting from the mapping and resting phenomena. We include domains and phenomena if these are missing. We complete the context diagram if the entire environment of the machine has been considered and added to the context diagram.
6. **Requirements refinement.** We consider the requirements specified earlier from the goals in the SI* models and refine these into requirements related to the context diagram. These requirements refer to the domains and phenomena in the context diagram. While on the abstraction level of the requirements derived from goals, we might state that a some data should be kept confidential. The same statement would have to be refined. For example, the data has to be kept confidential within a certain database, represented as a causal domain. We have to refine each requirement and check carefully if further requirements need to be added. This can be the case, because the context diagram might contain elements and relations that were not present in the SI* models.
7. **Create problem diagrams.** We create at least one problem diagram for each requirement defined in the previous step. The problem diagrams contain all elements the requirements refer to from the context diagram. It is also possible to add further domains to the problems diagrams. We consider the set of resulting domains from the mapping of SI* elements to elements from the problem frame notation. At

Table 1. Mapping of Concepts from SI* models to Problem Frames

SI* Element	PF Element
Role or Actor	Biddable Domain or Causal Domain
Goal	textual Requirement
Task	basis for at least one Phenomenon
Resource	Lexical Domain or Causal Domain
Means end(Goal, Task)	*Actor!Task* shared between *Machine* and *Actor* Biddable Domain
Means end(Goal, Resource)	*Resource* Lexical Domain constrained by *Requirement* which refines *Goal*
Means end(Task, Resource)	*Actor!Task* shared between *Actor* Biddable Domain and Machine of the problem diagram of the related Requirement. Machine!Task between Machine and the Lexical or Causal Domain representing the *Resource*.
De(Depender, Dependee, Dependum)	If *Dependum* is a Goal the Biddable Domain *Dependee* is constrainted by the Requirement which refines *Goal* If *Dependum* is a Task *Dependee!Task* If *Dependum* is a Resource *Dependee!Resource*
Dp(Depender, Dependee, Dependum)	*Depender!grantResource*, *Dependee!accessResource*

the completion of this step we check if all problem frame elements from the mapping have been used at least one problem diagram. If this is not the case we might have missed an problem diagram and a requirement. We have to correct this by include further requirements and problem diagrams for the missing problem frame elements.

8. **Problem-based conflict analysis.** This step focuses on a conflict analysis based upon the problem frame models. The aim of this analysis is to detect and resolve problems that can be identified on the abstraction level of problem frames. This includes technical analysis considering the differences between domains and phenomena. For example, if a phenomenon is directed in only one direction we have to ask ourselves if we really do not want a phenomenon in the other direction. We also have to consider if connections between domains are missing or if problem diagrams contradict each other. For example, if two problem diagrams add further domains and these are not compatible to one another.
9. **Create architecture.** The last step of our method is the creation of software architectures in UML from the problem frames as explained in [4].

4.2 Mapping

We present the rules to map SI* elements to Problem Frames elements in Tab. 1. Note that not all concepts and relations of SI* can be mapped to Problem Frames elements. We map a role or agent either to a biddable domain or to a causal domain. *Roles* or *Agents* map to a *Biddable Domain* if they represent human actors. If they represent entire businesses, hardware or even software, they are mapped to *Causal Domains*.

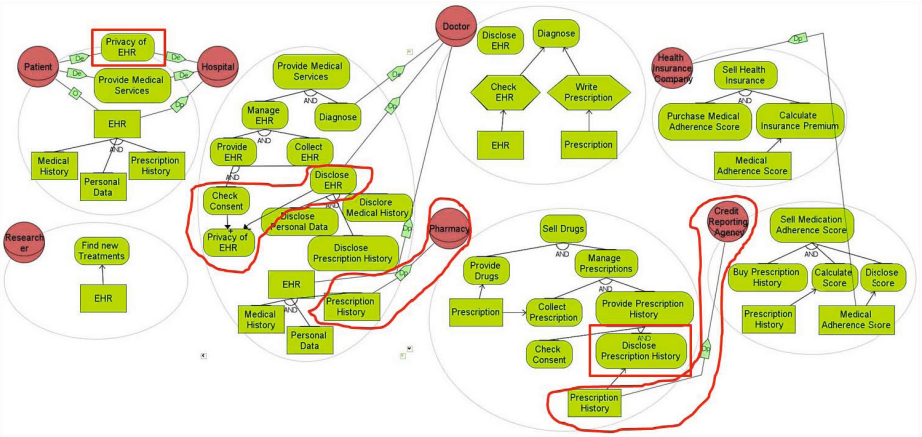
There is no representation of *Goals* in the Problem Frames notation. *Requirements* are means to fulfil goals [8] and are represented as text in the Problem Frame method, as well as graphical element in problem diagrams. *Tasks* in the SI* notation give rise to *Phenomenon* in the Problem Frames notation. The *Phenomenon* mapped to the *Task* will be under the control of the *Biddable Domain*, which has been mapped to the *Actor* providing the *Task*. A *Resource* in SI* is mapped to a *Lexical Domain* if it is just data. In the case it is hardware, software or any other mean with well defined behavior it is mapped to a *Causal Domain*.

Different mapping rules are applied to translate the relation *means_end* into Problem Frames elements based on the type of services (goal, task, resource) that are linked. If *means_end* is a relation between a *Goal* and a *Task* which is *provided* by an *Actor*, *Task* will be mapped to a *Phenomenon* under the control of the *Biddable Domain* representing *Actor*. The phenomenon belongs to the interface between the machine *Machine* and *Actor*. In addition, *Actor* is referred by the *Requirement* which refines the goal *Goal*. If *means_end* is a relation between a *Goal* and a *Resource*, *Resource* will be mapped to a *Lexical Domain* which is constrained by the *Requirement* that refines *Goal*. If *means_end* is a relation between a *Task* and a *Resource*, *Resource* is mapped to a lexical or causal domain. Additionally, there has to be Actor!Task between *Biddable Domain Actor* that represents the actor providing *Task* and the *Machine*. And Machine!Task between *Machine* and *Domain* representing *Resource* has to be added. To map the *delegation of execution* relation we use different rules according to the type of the Dependum. If the Dependum is a Goal, the Domain representing the *Dependee* Actor in the delegation of execution relation is *constrained* by the *Requirement* which is a refinement of Dependum *Goal*. If the Dependum is a Task or a Goal, the delegation of execution dependency is mapped to two Phenomenon under the control of the Biddable Domain representing the *Dependee* Actor. Furthermore, the *delegation of permission* relation having as Dependum a *Resource* is mapped to two *Phenomenon* under the control of the *Biddable Domain* representing the actors *Depender* and the *Dependee* because both of them have access to the *Lexical Domain* representing the *Resource*.

4.3 Consistency

We propose several consistency checks for our method:

- The *backtracking* from a problem frame diagram to a requirement to the goal has to be possible
  - The actors, which have provided a relation to a goal, have to appear in the problem diagram.
  - If a resource has a means-end relation to a goal, the resource has to appear in the problem diagram.
  - If a task has a means-end relation to a goal, at least one phenomenon related to that task has to be in the problem diagram
- If *changes* are made in the goal model these have to be also applied to the problem frame model



Legend: Circles represent roles (e.g Patient), ovals denote goals (e.g Provide Medical Services), exagons model tasks (e.g Write Prescription), while rectangles model resources (e.g EHR).

Fig. 4. SI* instantiation of the EHR scenario

5 Application of Our Method

Application scenario. To illustrate our method for combining goal- and problem-based requirements engineering methods, we use a scenario taken from the health care domain provided by the industrial partners of the EU project NESSoS [2]. It concerns the management of electronic healthcare records. The scenario focuses on providing medical services to patients, reading and updating their healthcare record and providing the results of examinations and treatment to authorized externals entities.

1. Instantiate SI* Models. Fig. 4 shows the SI* model for the management of electronic health records (EHRs). The model consists of seven actors: *Patient*, *Hospital*, *Doctor*, *Pharmacy*, *Credit Reporting Agency*, *Health Insurance Company* and *Researcher*. The *Patient owns* the resource Electronic Health Care Record (EHR) which consists of *Medical History*, *Prescription History*, and *Personal Data*. The *Patient* wants *Privacy of EHR* and *Provide Medical Services* to be guarantee and depends on the *Hospital* for these goals to be achieved. To achieve the goal *Provide Medical Services*, *Hospital* needs to *Manage EHR* and the *Diagnose* the *Patient*. The goal *Manage EHR* is decomposed in the subgoals *Provide EHR* and *Collect EHR*. *Provide EHR* is further decomposed into subgoals *Check Consent* and *Disclose EHR*. The goal *Check Consent* positively contributes to the achievement of the goal *Privacy of EHR*, while the goal *Disclose EHR* negatively affects the fulfillement of the goal. To satisfy the goal *Diagnose* the *Hospital* depends on the *Doctor*. The *Diagnose* goal is achieved by the tasks *Check EHR* and *Write Prescription*. The task *Check EHR* requires the resource *EHR* to be accomplished and thus the *Hospital* delegates to the *Doctor* both the permission of accessing *EHR* and of *Disclose EHR*. The *Pharmacy* is responsible for *Sell Drugs* to

[2] http://www.nessos-project.eu/

patients. The goal *Sell Drugs* consists of *Provide Drugs* to be delivered according to the *Prescription* received from the *Patient* and *Manage Prescriptions*. The goal *Manage Prescriptions* requires to *Collect Prescription* and *Provide Prescription History*. This goals requires to *Check Consent* and *Disclose Prescription History*. The *Credit Reporting Agency* wants to *Sell Medical Adherence Score* that indicates how well patients handle drug prescriptions. To this end, *Credit Reporting Agency Buy Prescription History* to *Calculate Score* and *Disclose Score* to third parties like *Health Insurance Companies*. *Health Insurance Company Purchase the Medical Adherence Score* to calculate the *Health Insurance Premium*. Finally, the information in the *EHR* is used by the *Researcher* to *Find new Treatments*.

Figure 4 shows a fragment of the SI* model in Fig. 4 where a conflict occurs between the main goal of the *Patient* and *Pharmacy*. The conflict arises between the main goal of *Patient*, *Privacy of EHR* and the *Pharmacy* goal *Disclose Prescription History* (wrapped by red rectangles). The *Patient* has delegated the execution of the goal *Privacy of EHR* to the *Hospital*. The fulfillment of the goal *Privacy of EHR* is negatively affected by the goal *Disclose EHR* which is decomposed in the sub-goals *Disclose Personal Data*, *Disclose Medical History* and *Disclose Medical History*. The *Hospital* has not delegated to the *Pharmacy* the permission of *Disclose Prescription History* but just the permission of accessing *Prescription History*. The *Pharmacy*, instead, grants access on *Prescription History* to *Credit Reporting Agency*. Thus, the *Pharmacy* goal *Disclose Prescription History* is in conflict with the goal *Privacy of EHR* of the *Patient*. The information about which goals have a negative impact on each other is domain knowledge. This cannot be derived automatically and has to be an input to the model. However, the detection of these goal conflicts is not possible in the Problem Frame notation, because stakeholder goals are not part of the notation. Thus, we can only detect these goal conflicts when using SI* (or another goal-based notation). We resolve the conflict by removing the goal *Provide Prescription History* to the *Pharmacy*.

2. Derive Requirements from Goals. We refined the goals from the SI* models into requirements. The goals depicted in Fig. 4 are marked in bold in the following requirements:

RQ1 Provide the means for the doctor to **Diagnose** patients
RQ4 Provide access to monitoring data for researchers, so that the researcher can **Find new Treatments**
RQ5 Provide means to enable doctors to **Asses Treatment** plans of other doctors

3. Map SI* to Problem Frame Elements. We provide a mapping of the SI* elements to the PF elements in Tab. 2.

4. Create Context Diagram. In Fig. 5 we present a context diagram of the electronic health system (***EHS***), the machine to be built. The EHS has connections to the lexical domains ***EHR***, the ***Privacy Policie***s of the system, and the ***System Log***s. The Privacy Policies state the privacy preferences of each patient, e.g., to which doctors the patient has given a consent to use her EHR. The EHS is connected to the ***Patient*** using the ***Browser Patient***, a ***Mobile Device***, which is further connected to ***Sensor***s that are in

Table 2. Mapping of SI* elements to PF elements

SI* Element	PF Element
Actors	
Patient	Biddable Domain *Patient*, because it represents a group of human actors
Hospital	The *EHS* machine offers no functionality to the hospital directly. Thus, we do not require a domain for it in the context diagram. The functionality is actually provided to the *Doctor*.
Doctor	Biddable Domain *Doctor*, because it represents a group of human actors
Researcher	Biddable Domain *Researcher*, because represents a group of human actors
Pharmacy	The *EHS* machine offers no functionality to the pharmacy directly. Thus, we do not require a domain for it in the context diagram. The functionality is actually provided to the *Pharmacist*.
Pharmacist	Biddable Domain *Pharmacist* is introduced, because these group of actors use the *EHS* machine.
Credit Reporting Agency	The *EHS* machine offers no functionality to the credit reporting agency. Thus, we do not require a domain for it in the context diagram. The prescription information is forwarded from *Pharmacists* with other means.
Health Insurance Company	The *EHS* machine offers no functionality to the health insurance company. Thus, we do not require a domain for it in the context diagram.
Tasks	
MonitorVitalSigns	A sequence of phenomena where the machine records the vital signs of the patient: *P!VitalSigns* and the according sequence of phenomena to the machine: *S!VitalSigns, MD!CreateEHR*
ProcessMonitoring-Data	The processing of the data is machine internal, but for processing the data has to be read first: *EHS!RequestEHR*
Notify about Treatment	A phenomenon that informs the nurse or doctor about a treatment to be applied and the corresponding sequence of phenomena: *EHS!sendEHR, M!sendEHR*
Record Appliance	A sequence of phenomena that sends an EHR with the treatment entry from the nurse / doctor to the machine: *N!sendEHR, D!sendEHR, BCP!sendEHR*
Check Health Data	A phenomenon that enables the doctor to request an EHR: *D!requestEHR, BCP!requestEHR*
Write Prescription	A phenomenon that enables the doctor to write an prescription: *D!send, BCP!sendEHR*
Provide Prescription	A phenomenon that enables the pharmacist to write an prescription: *P!provideDescription, BP!providePrescription,E!sendPrescription*
Collect Prescription	A sequence of phenomena that enables the pharmacist to collect a prescription: *P!collectPrescription, BP!collectPrescription,EHS!collectPrescription, E!sendPrescription,EHS!sendPrescription,BP!sendPrescription*
Resource	
Mobile Device	Causal Domain *Mobile Device*, because it is a device containing hard- and software. It is also a Connection Domain *Mobile Device*, because it connects the sensor and the EHS.
MonitoringData	Lexical Domain *Monitoring Data*, because it is data without a physical device.
Sensor	Causal Domain *Sensor*, because it is a device containing hard- and software.
Electronic Health Record (EHR)	Lexical Domain *EHR*, because it is data without a physical device.
Prescription	Lexical Domain *EHR*, because it is data without a physical device.

turn attached to the Patient. A ***Monitor*** or the ***Browser Care Provider***s connects the machine to the health care professionals, in our example ***Doctor***s.

The ***Researcher*** uses the ***Browser Researcher*** to access the ***Research Database Application***, which is in turn connected to the EHS. The hospital from the SI* diagram is not mapped to in the context diagram, because it has no direct connection to the machine. The reason is that the hospital delegates all goals to other actors. Hence, a direct connection to the machine is not necessary. The Problem Frame method demands that the connections between domains and the machine are made explicit. Hence, our model got enriched by several connection and causal domains, e.g., ***Browser Researcher***. These are not present in the SI* models.

5. Refine Requirements. We identified 19 preliminary functional requirements for the EHS, which were refined into 34 functional requirements and corresponding problem diagrams. For reasons of space, we focus on the following refined requirements for the remainder of the paper:

RQ1 Provide the means for the doctor to diagnose patients
- **RQ1.1** Store ***EHR***, which is created by care providers.
- **RQ1.2** Display ***EHR*** to care providers as needed
- **RQ1.3** Store and process vital signs of patients in ***EHR*** for care providers

RQ4 Provide access to monitoring data for researchers, so that the researcher can find new treatments
- **RQ4.1** Provide a functionality to request medical data for ***Researcher***s
- **RQ4.2** Release medical data to ***Researcher***s

6. Create Problem Diagrams. The problem diagram for RQ1.1 describes creating and storing of ***EHR***s (depicted in Fig. 6). The ***Patient*** is connected to a ***Sensor*** that reports the ***Patient***'s vital signs to the ***EHR Create & Store Machine*** using a ***Mobile Device***. The machine stores the ***EHR***. In addition, the ***Patient*** can use the ***Browser Patient*** to create an ***EHR***. ***Doctor***s and ***Nurse***s can use the ***Browser Care Providers*** to command the machine to create ***EHR***s.

The release of medical information to researches described in RQ4.2 and depicted in the problem diagram in Fig. 7. ***Researcher***s can use the ***Browser Researcher*** to request medical data from the ***Research Database Application***. This application requests the data in turn from the ***ReleaseMedicalDataMachine***, which releases it the ***Research Database Application***. The application sends the information to the browser, where it is shown to the ***Researcher***s.

The remaining problem diagrams are drawn in a similar manner.

7. Problem-Based Conflict Analysis. We proposed a pattern-based method for identifying laws [3] based upon software requirements specification. Using this method for our case study gives the result that data protection law, among others, is relevant. For example, the the German data protection act (BDSG). This law demands a detailed privacy analysis of a system. In earlier works we proposed the ProPAn method [2] for computer-aided privacy threat analysis. The method is based upon the problem frame

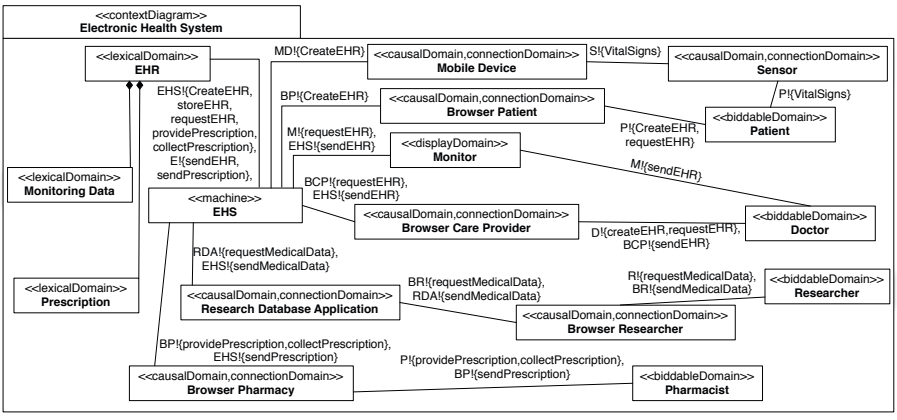


Fig. 5. Context Diagram

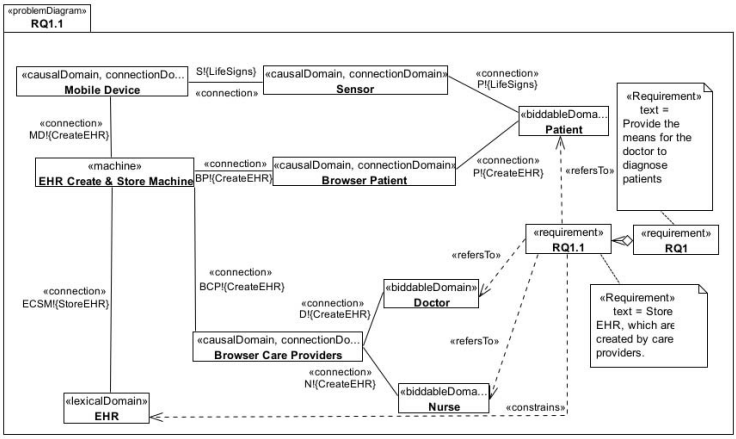


Fig. 6. Problem Diagram for Requirement RQ1.1

method and results in a set of requirements that have a privacy conflict. In our case study the results are that RQ4 has a conflict with other requirements, because the *Researcher*s gain access to personal information of the *Patient*s without an informed consent. We resolve this issue by anonymising the data transferred to the researchers. The ProPAn method only uses the problem frame approach and relies upon an analysis of the domains and their connections. Hence, the complete modelling of the machine and all the domains in the environment allows to trace personal information through the system. This complete tracing is difficult to achieved with SI*, because the system is only modelled in relevant artifacts. Hence, it might be possible to detect a privacy violation, if a related goal is modelled. To describe a tracing, however, is rather difficult to master in SI*, because the system is only partially modelled.

After changing the problem diagram of RQ4, we have to execute consistency checks on the other problem frame models and the SI* models. This also results in a Feedback loop to the goal model and we have to update the goal model accordingly. This results in the model shown in Fig. 8. In this model, we added goal the *protect personal data* of the new established actor *Legislator* (red part with the lightning). The *Hospital* has to comply to this law. Hence, we have to include the information that the data in the EHR have to be anonymised, before sending to the *Researcher*s (green part with the exclamation mark).

8. Create Architecture from Problem Diagrams. For space reasons, we refer to the works of Choppy et al. [4] for creating software architectures from problem diagrams.

6 Discussion

The integration of a goal-based and a problem-based requirements engineering method can be achieved either by integrating the concepts of a goal-based method into

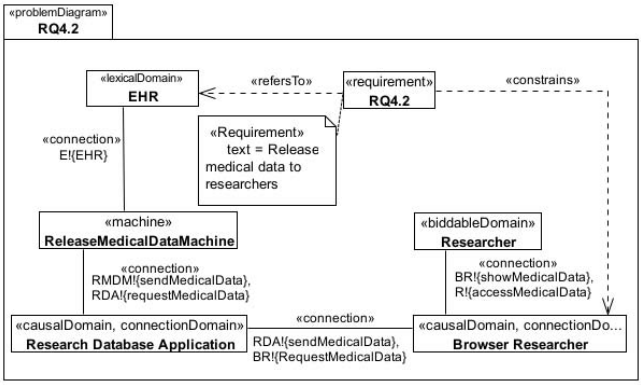


Fig. 7. Problem Diagram for Requirement RQ4.2

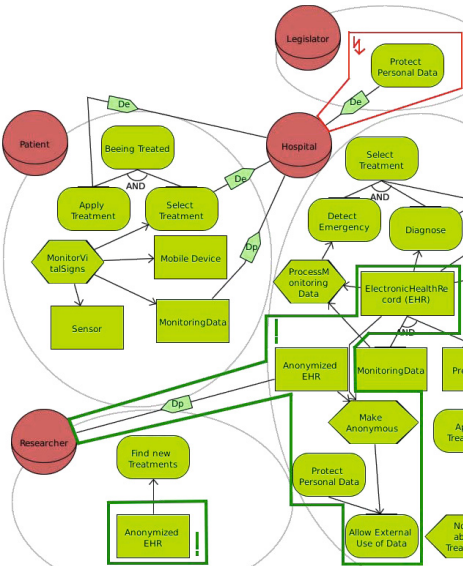


Fig. 8. Patient Monitoring - An example of a revised goal model due to an requirements conflict

a problem based (and vice versa) or by creating relations between a goal-based and a problem-based method. We did experiments with the integration of goals and more explicit stakeholder views into the Problem Frame approach. However, the results were difficult to process, because we discovered that the goals are on a higher abstraction level the Problem Frame method. For example, creating a goal domain and instantiating it with the goal *Earn Money* is feasible. However, each domain in the Problem Frame approach has to have relations to other domains, and phenomena towards other domains or the machine. We could not come up with meaningful phenomena between goals and Problem Frame domains. Moreover, the investigation of goal conflicts is very

difficult, when the relations between stakeholder goals are expressed in phenomena. Hence we abandoned the approach of integrating concepts.

Nevertheless, we agree that a intensive analysis of stakeholder goals is valuable for every software development problem. Thus, we decided to work on relations between the two requirements engineering methods. This has the drawback that two time consuming methods have to be executed. However, we recognized that it is important to analysis conflicts on different levels of abstractions. Goal-based methods have the advantage of only considering the technical aspects of the software if they have a direct relation to a stakeholder goal. In addition, goal-based methods only consider the specific parts of a software that has this relation. The problem-oriented world works on a lower abstraction level and tries to describe a complete context of the machine. This has the advantage of being able to derive software architectures from this context description.

7 Conclusions

In this paper we have presented a method to integrate goal- and problem-based requirements engineering methods. We presented a particular method for the integration of the SI* and Problem Frames methods. This method provides the ability for software engineers to analyze and resolve goal conflicts among stakeholders on a high abstraction level and to transfer the resulting goal models to problem models that focus on the particular software design problem.

Our method offers the following main benefits:

- Combining the SI* and Problem Frames methods without changing the existing methods
- Systematic identification and solving of goal conflicts
- Beginning the Problem Frames method with already resolved goal conflicts
- Using the advantages of goal- and problem-based requirements engineering methods

In the future, we will use or method for further goal-based methods, e.g., KAOS [20]. We will also look into the integration of security- and risk-based requirements engineering methods, e.g., CORAS [16]. We will also investigate the applicability of graph model transformation techniques to semi-automate some steps of our method like the mapping from SI* to Problem Frames elements.

References

1. Alebrahim, A., Hatebur, D., Heisel, M.: A method to derive software architectures from quality requirements. In: Thu, T.D., Leung, K. (eds.) Proceedings of the 18th Asia-Pacific Software Engineering Conference (APSEC), pp. 322–330. IEEE Computer Society Press (2011)
2. Beckers, K., Faßbender, S., Heisel, M., Meis, R.: A problem-based approach for computer aided privacy threat identification. In: Privacy Forum. LNCS. Springer (2012) (to appear)
3. Beckers, K., Faßbender, S., Küster, J.-C., Schmidt, H.: A pattern-based method for identifying and analyzing laws. In: Regnell, B., Damian, D. (eds.) REFSQ 2011. LNCS, vol. 7195, pp. 256–262. Springer, Heidelberg (2012)

4. Choppy, C., Hatebur, D., Heisel, M.: Systematic architectural design based on problem patterns. In: Avgeriou, P., Grundy, J., Hall, J., Lago, P., Mistrik, I. (eds.) Relating Software Requirements and Architectures, ch. 9, pp. 133–159. Springer (2011)
5. Classen, A., Heymans, P., Laney, R., Nuseibeh, B., Tun, T.T.: On the structure of problem variability: From feature diagrams to problem frames. In: Proceedings of International Workshop on Variability Modeling of Software-Intensive Systems, Limerick, Ireland, pp. 109–118 (January 2007)
6. Côté, I.: A Systematic Approach to Software Evolution. Deutscher Wissenschafts-Verlag (DWV) Baden-Baden (2012) (to appear)
7. Côté, I., Hatebur, D., Heisel, M., Schmidt, H.: UML4PF – a tool for problem-oriented requirements analysis. In: Proceedings of the International Conference on Requirements Engineering (RE), pp. 349–350. IEEE Computer Society Press (2011)
8. Fabian, B., Gürses, S., Heisel, M., Santen, T., Schmidt, H.: A comparison of security requirements engineering methods. Requirements Engineering – Special Issue on Security Requirements Engineering 15(1), 7–40 (2010)
9. Hatebur, D.: Pattern and Component-based Development of Dependable Systems. Deutscher Wissenschafts-Verlag (DWV) Baden-Baden (September 2012)
10. Hatebur, D., Heisel, M.: A foundation for requirements analysis of dependable software. In: Buth, B., Rabe, G., Seyfarth, T. (eds.) SAFECOMP 2009. LNCS, vol. 5775, pp. 311–325. Springer, Heidelberg (2009)
11. Hatebur, D., Heisel, M.: A UML profile for requirements analysis of dependable software. In: SAFECOMP, pp. 317–331 (2010)
12. Jackson, M.: Problem Frames. Analyzing and structuring software development problems. Addison-Wesley (2001)
13. Jackson, M., Zave, P.: Deriving specifications from requirements: an example. In: Proceedings 17th Int. Conf. on Software Engineering, pp. 15–24. ACM Press, Seattle (1995)
14. Letier, E., van Lamsweerde, A.: Deriving operational software specifications from system goals. SIGSOFT Softw. Eng. Notes 27(6), 119–128 (2002)
15. Liu, L., Jin, Z.: Integrating goals and problem frames in requirements analysis. In: 14th IEEE International Conference Requirements Engineering, pp. 349–350 (2006)
16. Lund, M.S., Solhaug, B., Stølen, K.: Model-Driven Risk Analysis: The CORAS Approach, 1st edn. Springer (2010)
17. Massacci, F., Mylopoulos, J., Paci, F., Tun, T., Yu, Y.: An extended ontology for security requirements. In: Advanced Information Systems Engineering Workshops, vol. 83, pp. 622–636 (June 2011)
18. Massacci, F., Mylopoulos, J., Zannone, N.: Security requirements engineering: The SI* modeling language and the secure tropos methodology. In: Ras, Z.W., Tsay, L.-S. (eds.) Advances in Intelligent Information Systems. SCI, vol. 265, pp. 147–174. Springer, Heidelberg (2010)
19. Supakkul, S., Chung, L.: Extending problem frames to deal with stakeholder problems: An agent- and goal-oriented approach. In: Jacobson Jr., M.J., Rijmen, V., Safavi-Naini, R. (eds.) SAC 2009. LNCS, vol. 5867, pp. 389–394. Springer, Heidelberg (2009)
20. van Lamsweerde, A.: Requirements Engineering: From System Goals to UML Models to Software Specifications, 1st edn. John Wiley & Sons (2009)

GPRS Security for Smart Meters

Martin Gilje Jaatun[1], Inger Anne Tøndel[1], and Geir M. Køien[2]

[1] Department of Software Engineering, Safety and Security
SINTEF ICT
NO-7465 Trondheim, Norway
{martin.g.jaatun,inger.a.tondel}@sintef.no
http://www.sintef.no/ses

[2] Department of Information and Communication Technology (ICT)
Faculty of Engineering and Science
University of Agder
NO-4879 Grimstad, Norway
geir.m.koien@uia.no
http://www.uia.no/en

Abstract. Many Smart Grid installations rely on General Packet Radio Service (GPRS) for wireless communication in Advanced Metering Infrastructures (AMI). In this paper we describe security functions available in GPRS, explaining authentication and encryption options, and evaluate how suitable it is for use in a Smart Grid environment. We conclude that suitability of GPRS depends on the chosen authentication and encryption functions, and on selecting a reliable and trustworthy mobile network operator.

Keywords: Security, GPRS, Smartgrid, AMI, Smart Metering.

1 Introduction

Smart Meters in an Advanced Metering Infrastructure (AMI) represent perhaps the most visible aspect of the Smart Grid [1]. Smart Meters are placed in every home, providing real-time two-way communication with the electricity provider, or Distribution Service Operator (DSO).

AMIs allow the automatic collection of power consumption data, and thereby more granular pricing schemes. However, AMIs also allow DSOs to have better overview of the current state of power delivery and increased opportunities for control. Smart meters can send status messages and alarms to the DSO, they may also have breaker functionality, allowing power to be turned off remotely. Furthermore, meters may be connected to equipment in the homes that allows automatic adjustment of power demand based on the current price level.

Smart Meters use a variety of communications means towards DSOs, but General Packet Radio Service (GPRS) is a popular choice in many rural settings [2]. GPRS was introduced as a faster data transfer service for GSM Mobile Stations, offering a packet-based data service according to the Internet Protocol (IP). It introduces a number of new network elements, most notably the Serving

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 195–207, 2013.

GPRS Support Node (SGSN) which connects to the Mobile Station via the current active Base Transceiver Station (BTS), and the Gateway GPRS Support Node (GGSN) at the Mobile Station's home operator, which connects to the public internet. Using GPRS for AMI implies equipping smart meters with a UICC/USIM or SIM subscriber module, and a communication terminal that communicates by means of GPRS.

Security is important in AMI systems. Smart meters will need to communicate personal information; power delivery to meters can be remotely controlled; and status updates, software updates and configuration parameters can all be considered confidential. Thus, DSOs need to know how communication is secured. Security can be used as a differentiating factor when deciding on which communication technology to use. It is however likely that other considerations, such as the area of coverage of the telecommunication network and how densely the area is populated [3], will be given more weight. Still, knowledge of the protection offered by the alternative communication technologies is important to decide what additional security measures are needed. Only with such knowledge can DSOs end up with AMI solutions that are both cost-effective and offer adequate security.

This paper explains the security offered by GPRS, when used for AMI. Section 2 gives an overview of threats towards AMI systems to provide a basis for understanding the main security needs for AMI. Section 3 provides an overview of the security functionality of GPRS. Section 4 summarises the security offered by GPRS based on the security needs of AMI explained in Section 2 and provides recommendations for DSOs that plan on using GPRS for communication with their smart meters. Section 5 concludes the paper.

2 Background

The increased connectivity and the new trust models that come with the introduction of AMI lead to new threats and a pressing need for DSOs to deal with information security and consumer privacy. The context in which we are considering GPRS communication is illustrated in Fig. 1. Tøndel et al. [1] identified threats towards AMI systems on the interface between a smart meter and the Head End System (HES) at the DSO. Threats include the following (threats are numbered in the same way as in the paper by Tøndel et al.):

- Fake identities: Someone takes the identity of a meter (T2) or the identity of the HES (T1)
- Tamper with communication: Someone tampers with the communication sent between the HES and meters (T5)
- Repudiation: Meter denies receiving a message (T10) or the sending of a message (T11)
- Eavesdropping: Someone eavesdrops on communication between a meter and the HES (T13)
- Denial of service: Communication is hampered due to a denial of service attack on the HES (T18), an attack on one or more meters (T19) or a failure on the communication link (T20)

For DSOs it is important to understand the extent to which GPRS and other candidate communication technologies are vulnerable to such attacks or help prevent failures of the above listed types. Also it is important to become aware of the configuration choices available that can increase or lower the achieved level of security.

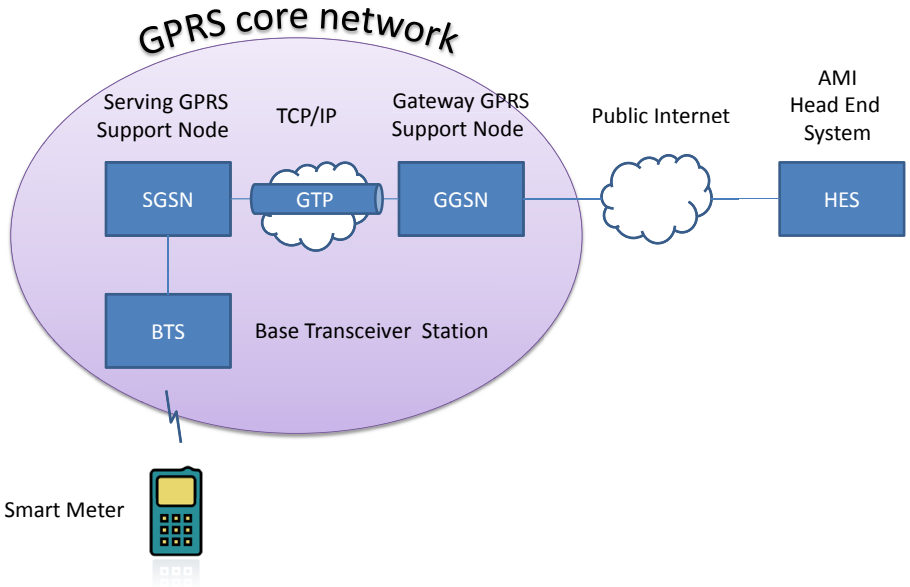


Fig. 1. GPRS communication for AMI

The ability to use fake identities is related to the authentication functionality of GPRS. The feasibility of eavesdropping and tampering is dependent on the degree of protection of the communication. The risk of repudiation problems are dependent on the degree to which actions can be proved. The risk of denial of service has to do with the capacity of the communication media and the equipment, as well as its robustness. In the following, we will describe available security mechanisms in GPRS in order to evaluate how these threats are handled.

3 GPRS Security Functions

GPRS offers security functionality on the interface between the mobile terminal and the GPRS core network. This section describes that functionality, including security functionality that can be achieved by use of UICC/USIM instead of SIM. It also provides a brief introduction to security in the GPRS core network, and means to protect information when sent from GPRS core to the DSO via the public internet.

3.1 Authentication and Key Agreement

In GPRS, mobile users are authenticated towards the network. The native 2G authentication and key agreement (AKA) protocol used in the packet-switched GPRS system is the same as the one used in the circuit-switched GSM system, but the challenge-response is performed by the Serving GPRS Support Node (SGSN) rather than the Visitor Location Register/Mobile Switching Center (VLR/MSC). We use the name VLR/SGSN to denote cases where there is no distinction between the GSM and GPRS cases. The authentication protocol is known as GSM-AKA, using the interface[1] functions A3 and A8 (TS 43.020 [4]). The actual implementation is operator dependent, but the GSM Association have several example algorithms available.

The GSM-AKA protocol is a two-stage protocol, as illustrated in Fig. 2:

- Request from VLR/SGSN to HLR/SGSN and forwarding of triplet (authentication set) from HLR/AuC to the VLR/SGSN (steps ① and ②)
- Network-initiated challenge-response (VLR/SGSN - SIM) (steps ③ and ④)

Table 1. Elements used in the GSM-AKA Authentication and Key Agreement protocol

K_c	Secret calculated by SIM and HLR/AuC, sent to VLR/SGSN
K_i	Secret key shared between SIM and AuC
RAND	Random challenge selected by VLR/SGSN
RES	Result of challenge calculated by SIM and sent to VLR/SGSN
XRES	Expected result of challenge calculated by HLR/AuC, sent to VLR/SGSN

The tamper resistant SIM card includes the permanent subscriber identity (International Mobile Subscriber Identity – IMSI), the per-subscriber authentication secret K_i and the AKA algorithms (A3/A8). The IMSI and K_i are one-to-one associated, and the K_i is a pre-shared secret. The IMSI will be known to multiple nodes and networks, but the K_i is only known to the SIM and the HLR/AuC.

The HLR/AuC will produce triplets $(RAND, SRES, K_C)$ and forward these to the VLR/SGSN upon request. The $RAND$ is 128-bit wide and is the (pseudo) random challenge while the 32-bit $SRES$ is the signed response. The A3/A8 are one-way functions (Fig. 3), ideally making it computationally infeasible for an intruder to determine the K_i and the K_C upon witnessing the plaintext exchange of the challenge-response procedure[2]. The VLR/SGSN receives the triplet and

[1] The various Ax, fy functions defined in GSM and UMTS (A3, A5, f1, f2, ...) are only standardised at the interface level; i.e., expected input and output is defined, but the actual implementation is left to each operator. However, many operators employ one of the example implementations documented by The GSM Association and 3GPP, as explained later in this paper.

[2] Note that it must be computationally infeasible to recover K_i also when the attacker can observe a large number of challenge-response exchanges, ideally also for chosen-plaintext attacks. Even if an attacker should have access to the full triplet (including $SRES$ and K_c) it should not be possible to recover K_i.

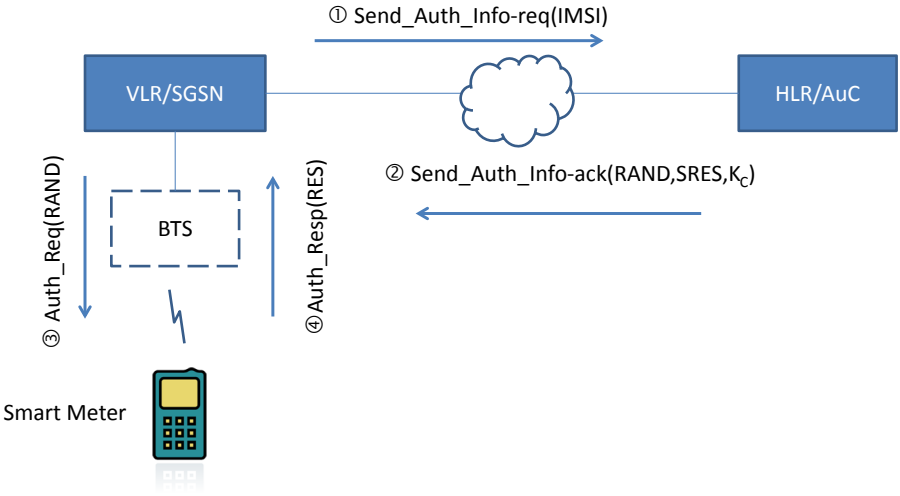


Fig. 2. The GSM-AKA Authentication and Key Agreement protocol

challenges the SIM by forwarding the $RAND$. The SIM will compute the $SRES$ and K_C, and the $SRES$ will be forwarded as the response to the VLR/SGSN. The encryption key will be forwarded to the MS from the SIM upon request. Provided that the $SRES$ from the HLR/AuC matches the result from the SIM, the authentication is considered successful. The outcome also includes the SIM and the VLR/SGSN sharing the K_C encryption key. The ciphering key used is the 64-bit K_C key for both GSM and GPRS.

It should be noted that there exist some rather weak A3/A8 implementations, and specifically the original "example" algorithm known as COMP128 (completely broken [5,6]). Also note that GSM-AKA only generates a 64-bit session key, which clearly is not future proof considering the continual improvements in processing speeds for exhaustive key search applications. For standard GSM-AKA, we recommend the use of the GSM-Milenage [7] implementation, which uses the UMTS Milenage AKA functions and a set of conversion functions to implement the A3/A8 functions.

3.2 UMTS Authentication and Key Agreement

It is possible to run the UMTS Authentication and Key Agreement (UMTS-AKA) [8,9] protocol over the "GSM Edge Radio Access Network" (GERAN) if the subscriber has a UICC/USIM module instead of the old GSM SIM. The UMTS AKA protocol is defined in 3GPP TS 33.102 [10] and the functions are shown in Fig. 4. All the interface functions f1-f5 take the secret key K and a random challenge RAND as input, while f1 also takes a sequence number SQN

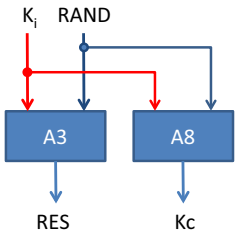


Fig. 3. Using the GSM A3/A8 functions to perform GSM-AKA calculations

and an Authentication Management Field AMF as input. The functions produce the following outputs:

MAC-A Message Authentication Code
XRES Expected challenge result
CK Cipher Key for subsequent message encryption
IK Integrity Key
AK Anonymity Key (for obfuscating the sequence number in case this exposes the identity of the client)

These values are computed by the Authentication Center (AuC), and further

$$\text{AUTN} = \text{SQN} \oplus \text{AK}||\text{AMF}||\text{MAC-A}$$

$$\text{AV} = \text{RAND}||\text{XRES}||\text{CK}||\text{IK}||\text{AUTN}$$

The Authentication Vector (AV) can be considered the UMTS equivalent to the GSM triplet. See Table 2 for a complete overview of the elements used in UMTS-AKA.

The Authentication Vector (AV) is roughly equivalent to the GSM triplet. The AK is used to mask the sequence number. The challenge from the SGSN includes the (RAND,AUTN) tuple. The UICC/USIM runs the calculations depicted in Fig. 4, and returns the result of f2 (RES). The SGSN compares the RES from the MS with the XRES from the AV, and if they match, the MS is authenticated.

3GPP has provided the Milenage algorithm set [11] as an example implementation of the f1-f5 functions. The Milenage algorithm set includes a cryptographic core, Rijndael (i.e., the algorithm implementing the Advanced Encryption Standard (AES)), and a set of constants and parameters to produce independent output from all the UMTS AKA functions.

3.3 GPRS Encryption

With GPRS, the data and signaling traffic is encrypted over the wireless interface and all the way from the mobile station to the GPRS core. The ciphering algorithms used in GPRS are known as the GEA family of algorithms (GEA,

Table 2. Elements used in the UMTS-AKA protocol

Acronym	Explanation
AK	48 bit Anonymity Key
AMF	16 bit Authentication Management Field
AUTN	Authentication values computed by AuC
AV	Authentication vector computed by AuC, sent to SGSN (equivalent to the GSM triplet)
CK	128 bit Cipher key, for encryption of data (confidentiality key)
IK	128 bit Integrity key
K	128 bit Secret key pre-shared between USIM and AuC
MAC-A	64 bit Message Authentication Code, "signature" to authenticate the challenge
RAND	128 bit Random challenge selected by VLR/SGSN
RES	32-128 bit Result of challenge calculated by USIM and sent to SGSN (default 64 bit)
SQN	48 bit Sequence Number
XRES	32-128 bit Expected result of challenge calculated by AuC, sent to SGSN (default 64 bit)

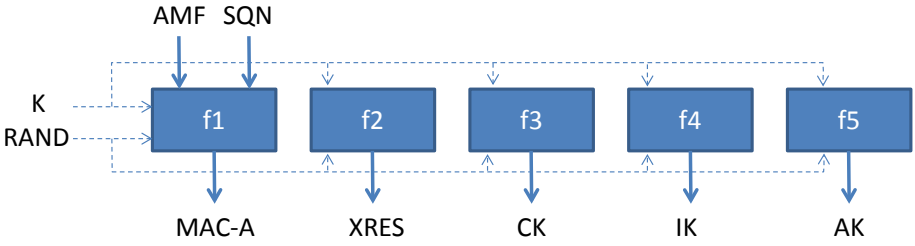


Fig. 4. UMTS Authentication and Key Agreement (AKA) functions

GEA2, GEA3, GEA4). The GEA interface is defined in TS 41.061 [12]. We note again that in 2G GPRS, the ciphering terminates in the core network (CN) in the SGSN, while in GSM the ciphering terminates in the BTS. The GSM equivalent ciphering algorithms are known as the A5 family of algorithms (A5/1, A5/2, A5/3, A5/4), where the A5/2 algorithm has been deprecated and where the A5/1 algorithm is now showing its age. The ciphering key used is the 64-bit K_C key for both GSM and GPRS, with the exception of GEA4 (and A5/4), which uses a 128 bit key.

The original GPRS algorithms (GEA and GEA2) are 64-bit designs and they are in fact similar, but the GEA algorithm is modified to only use 54 significant bits. This dates back to cold war export regulations, and the relaxation of export conditions allowed the GEA2 algorithm to regain the 10 lost bits.

The GEA3 algorithm is defined by the Keystream Generator Core (KGCORE) block cipher mode of operation, which in turn uses the the KASUMI[3] block cipher. The KGCORE primitive is based on a 128-bit key internally, and the interface achieves this by using the K_C key twice. GPRS can also use the GEA4 encryption algorithm with a 128 bit key called K_{C128}. GEA4 also uses KGCORE, but there the K_{C128} key is used directly. In order to use GEA4 (or A5/4 in GSM) the subscriber must use a UICC/USIM and run the UMTS AKA protocol.

We note that while a 3G/UMTS identity module (UICC/USIM) may be used in a 2G device (GSM/GPRS), the cipher function must inevitably be compatible with the over-the-air interface. This means that the UMTS key pair (CK,IK) must be converted into the the 64-bit K_C key (or the 128 bit K_{C128} key for GEA4).

The selection of encryption algorithm takes place by the mobile station indicating which version(s) of GEA it supports, and then the SGSN decides which version will be used [13]. If they do not support a common algorithm, the connection may be released. It is also an option that the SGSN decides on not encrypting the data.

There are no built-in security functions in GPRS that offer confidentiality or integrity protection beyond the SGSN[4]; specifically, all data sent on the public internet interface (Gi) from the GGSN is sent in plaintext. In fact, integrity protection is not provided at at all for GERAN. However, it is possible to configure so-called access point names (APNs) at the GGSN, and it is possible to use this mechanism to set up a VPN tunnel from the G_i-interface to an external network. To some extent such solutions[5] are already in commercial use.

3.4 Core Network Protection

The access security in GPRS terminates at the SGSN. Data traffic is forwarded between the SGSN and the GGSN by means of the GPRS Tunneling Protocol (GTP). There are various versions of GTP and there is a distinction between user plane (GTP-U) and control plane (GTP-C). Whereas the access may be by means of GERAN, the core network may still be 3G or 4G compliant. The GTP protocol (any version) does not by itself provide any security. There is not a strict requirement in 3GPP to cryptographically protect the GTP protocol, but it certainly is recommended [10]. Specifically, the 3GPP have profiled IPsec for use within 3GPP systems (TS 33.210 [14]) in what is known as the Network Domain Security (NDS) area. The NDS/IP [14] was originally developed for inter-operator use, but it increasingly being used for all IP-based interfaces in 3GPP system. We strongly recommend that NDS/IP be used for protection of GTP.

[3] Despite appearances, KASUMI is a proper noun ("mist" in Japanese), not an acronym.

[4] However, as will be described below, from SGSN to GGSN data is sent over GTP, which may offer additional protection options.

[5] See for instance `http://www.telenorfusion.no/makeit/communicationapis/mobiledataaccess/mdatechnicaldetails.jsp`

3.5 GPRS Closed User Groups

It is possible to set up closed user groups in the cellular networks [15], but as there is no appreciable security provided this can at best serve as a defence-in-depth measure. Furthermore, the service is really tailored for circuit-switched calls and is therefore largely irrelevant.

4 Discussion

In Section 2, we listed a set of risks relevant to the communication between a smart meter and the HES at DSOs. On this interface we claimed that DSOs need to be aware of the availability and strength of authentication mechanisms, the extent to which communication is protected, the degree to which actions can be proved, and the capacity and robustness of the communication media and the equipment used. In this section we sum up what support GPRS offers in this respect. We also identify technology and configuration choices that DSOs should make in order to improve the security of a GPRS communication solution used for AMI. The findings are summarised in Table 3.

Authentication of smart meter terminals to the communication network is performed by GPRS. The strength of this authentication is dependent on the authentication mechanism used, where the main alternatives, as explained above, are GSM-AKA and UMTS-AKA. Implementations of GSM-AKA algorithms (A3/A8) are operator dependent, and it should be noted that some of the available implementations are rather weak. Using the GSM-Milenage implementations of the A3 and A8 algorithms is thus recommended. For DSOs, the strength of the algorithm employed could be used as a differentiating factor among GPRS providers. DSOs should however consider equipping new smart meters with UICC/USIM. This will allow using UMTS-AKA for authentication, and also allow for additional security improvements over the use of SIM.

In GPRS, there is no mechanism for authenticating the communication network to the smart meter. This implies that it is possible to trick the meter into using a fake base transceiver station. According to Mitchell [16], it is no longer unfeasible for attackers with modest budgets to acquire a fake base transceiver station, and indeed news reports stated three years ago that a fake GSM base station could be built for as little as £1000 [17] - this sum is likely to be even lower today. However, if a meter is equipped with a UICC/USIM, two-way authentication can be performed (through the use of the MAC-A "signature" to verify the authentication challenge), and the fake base transceiver station threat is effectively mitigated.

GPRS communication that is sent on the wireless link is encrypted, but the encryption terminates in the core network. The strength of the encryption is dependent on the algorithm used, and it is also possible for the SGSN to specify that the traffic shall not be encrypted. Standard GPRS with a traditional GSM SIM can only support 64-bit encryption, making GEA, GEA2 and GEA3 the available algorithms. With a USIM, GEA4 is also an option. We recommend that smart meters use GEA3 or above for encryption. The smart meters thus need to

support one of these algorithms, and only advertise these during authentication. The SGSN also needs to support the algorithm, but as smart meters are not likely to move around, the support of one of these algorithms can be checked beforehand when selecting an operator. In other words, it is unlikely that a smart meter would need conventional roaming capability at all, always only connecting directly to the home mobile network operator. However, it should be pointed out that availability considerations might dictate that a smart meter should be able to use alternative providers if the home network provider is unavailable [2].

Generally, there is no default security in the protocols used in the GPRS core network. Operators may however choose to implement NDS/IP for protection of the GTP traffic; this is something we would recommend. DSOs should consider the trustworthiness and competence of the operator, as the operator's perimeter protection as well as its staff and internal routines are essential for the security of the communication. Selection of an operator implies that the DSO has to trust that operator's internal network.

From the GPRS core network the communication is sent unencrypted on the Internet on its way to the HES, unless additional protection measures are applied. Operators often offer VPN solutions, so that communication is sent in a secured tunnel from the GPRS core network to the network of the DSO. If the operator offers such solutions, we recommend that DSOs use them in order to protect the data sent to and from the smart meters.

Note that GPRS offers no mechanisms to prove actions (i.e, there is no non-repudiation functionality). If this is needed, then mechanisms must be added on top of GPRS. There is also no specific protection against denial of service attacks, and a fake base transceiver station could potentially effect a denial-of-service attack against a meter by tying it up indefinitely (this attack can be prevented by using UICC/USIM instead of SIM). Using wireless protocols is always challenging with respect to denial of service, as wireless media can easily be jammed. That being said, we are not aware of any particular weaknesses (compared to similar wireless technologies) in the GPRS technology when it comes to capacity and robustness of the communication media and equipment used. This may however depend on the operator.

As AMI systems will need to communicate confidential information; such as detailed meter consumption, configuration changes, software updates, and breaker commands; protection of the communication is essential. It is however important to note that AMI protection is not solely dependent on the security of the communication protocols. If GPRS is not able to offer strong enough protection, then additional security can be added on the application layer. This includes encryption, integrity protection and non-repudiation mechanisms. For DSOs that utilise several different technologies for meter communication, such application layer protection may be needed anyway because of different levels of security in the communication technologies used. It is however important that DSOs are aware of the protection offered, so that they can end up with a solution that is cost-effective while still offering adequate security.

Table 3. AMI threats evaluated based on using GPRS communication

Description	Threat#	Evaluation
Fake identities		
Fake HES	T1	GPRS with normal SIM will not prevent fake base transceiver stations (SGSN), and thus no protection against a fake HES. GPRS with UICC/USIM can perform mutual authentication toward a base transceiver station, and prevent introduction of fake HES in GPRS scope (but additional protection may be needed beyond the GGSN)
Fake meter	T2	GPRS authentication mechanisms will prevent fake meters from connecting to the network
Tamper with communication		
	T5	GPRS offers reasonable protection against tampering with GEA3 encryption (GEA4 offers good protection) (but additional protection may be needed beyond the GGSN)
Repudiation		
Reception	T10	GPRS offers no non-repudiation mechanisms
Sending	T11	GPRS offers no non-repudiation mechanisms
Eavesdropping		
	T13	GPRS offers reasonable protection against eavesdropping with GEA3 encryption (GEA4 offers good protection) (but additional protection may be needed beyond the GGSN)
Denial of service		
Attack on HES	T18	GPRS authentication mechanisms will prevent fake meters or other nodes from connecting with the HES
Attack on meter	T19	GPRS with normal SIM will not prevent fake base transceiver stations, which could be used to tie up a meter indefinitely. GPRS with UICC/USIM can perform mutual authentication of base transceiver station, ensuring that fake base transceiver stations are dropped immediately
Failure of communication link	T20	GPRS has no special protection against jamming, but neither does most commercially available wireless technology

5 Conclusion and Further Work

This paper has provided an overview of the security of GPRS related to its use for AMI. From a security point of view, we recommend the use of UICC/USIM instead of the SIM as this offers stronger and more extensive authentication and encryption. If using SIM, we recommend the use of the GEA3 encryption algorithm. We also recommend that DSOs use security and trustworthiness as a differentiating factor when choosing among mobile network operators.

There may be a need for security on top of GPRS. In particular, the communication between the GPRS core network and the DSO needs to be protected, either by using a VPN solution offered by the operator, or by implementing additional security functionality in higher layers. In a longer term perspective, additional security mechanisms may also contribute to future-proofing smart grid communication solutions that are likely to have a longer life span than their mobile telephony counterparts. Mobile communications technologies are constantly evolving, and it remains to be seen how e.g. LTE will influence AMI.

Acknowledgment. The research reported in this paper has been supported by the Telenor-SINTEF collaboration project, Smart Grid initiative.

Abbreviations

The world of mobile communications is drenched in three- and four-letter abbreviations and acronyms. To ease the reader's burden, we provide a list of the abbreviations used in this document here.

3GPP 3rd Generation Partnership Project
AKA Authentication and Key Agreement
APN Access Point Name
AuC Authentication Centre
AUTN AUThenticatioN value
BTS Base Transceiver Station (known as the Base Station in GSM)
GEA GPRS Encryption Algorithm (1-4)
GERAN GSM EDGE Radio Access Network
GGSN Gateway GPRS Support Node
GPRS General Packet Radio System
GSM Global System for Mobile Communications, originally Groupe Spécial Mobile
GTP GPRS Tunneling Protocol
HES Head End System
HLR Home Location Register
IMSI International Mobile Subscriber Identity
KGCORE Keystream Generator Core
LTE Long-term Evolution (also known as 4G LTE)
MS Mobile Station (Mobile telephone handset, or – in our case – a stationary Smart Meter)
MSC Mobile Switching Center
NDS Network Domain Security
SIM Subscriber Identity Module
SGSN Serving GPRS Support Node
TS Technical Specification
UICC Universal Integrated Circuit Card
UMTS Universal Mobile Telephone System
USIM Universal Subscriber Identity Module
VLR Visitor Location Register

References

1. Tøndel, I.A., Jaatun, M.G., Line, M.B.: Threat Modeling of AMI. In: Proceedings of the 7th International Conference on Critical Information Infrastructures Security, CRITIS 2012 (2012)
2. Telenor: Smart metering white paper – Best practices recommendation by Telenor Connexion (2013)
3. Gungor, V., Sahin, D., Kocak, T., Ergut, S., Buccella, C., Cecati, C., Hancke, G.: Smart grid and smart homes: Key players and pilot projects. IEEE Industrial Electronics Magazine 6(4), 18–34 (2012)
4. 3GPP: 3rd Generation Partnership Project; Technical Specification Group Services and system Aspects; Security related network functions (Release 12), 3GPP TS 43.020 V12.0.0 (2013)
5. SDA: Smartcard Developer Association Clones Digital GSM Cellphones (1998), http://www.isaac.cs.berkeley.edu/isaac/gsm-press.html
6. Rao, J.R., Rohatgi, P., Scherzer, H., Tinguely, S.: Partitioning attacks: or how to rapidly clone some gsm cards. In: Proceedings of 2002 IEEE Symposium on Security and Privacy, pp. 31–41 (2002)
7. 3GPP: 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3G Security; Specification of the GSM-MILENAGE Algorithms: An example algorithm set for the GSM authentication and key generation functions A3 and A8 (Release 8), 3GPP TS 55.205 V8.0.0 (2008)
8. Køien, G.M.: An introduction to access security in UMTS. IEEE Wireless Communications 11(1), 8–18 (2004)
9. Niemi, V., Nyberg, K.: UMTS security. John Wiley & Sons (2003)
10. 3GPP: Technical Specification 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3G Security; Security architecture (Release 11), 3GPP TS 33.102 V11.0.0 (2012)
11. 3GPP: 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3G Security; Specification of the MILENAGE Algorithm Set: An example algorithm set for the 3GPP authentication and key generation functions f1, f1*, f2, f3, f4, f5 and f5*; Document 1: General (Release 1999), 3GPP TS 35.205 V3.0.0 (2001)
12. 3GPP: 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); GPRS ciphering algorithm requirements (Release 4), 3GPP TS 41.061 V4.0.0 (2002)
13. Xenakis, C.: Security measures and weaknesses of the gprs security architecture. International Journal of Network Security 6(2), 158–169 (2008)
14. 3GPP: Technical Specification 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3G security; Network Domain Security (NDS); IP network layer security (Release 12), 3GPP TS 33.210 V12.0.0 (2012)
15. 3GPP: 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Closed User Group (CUG) Supplementary Services - Stage 1 (Release 9), 3GPP TS 22.085 V9.0.0 (2009)
16. Mitchell, C.J.: The security of the GSM air interface protocol (2001), http://www.ma.rhul.ac.uk/static/techrep/2001/RHUL-MA-2001-3.pdf
17. Stanley, N.: Mobile Phone Hacking for £1000. ComputerWeekly (2010), http://www.computerweekly.com/blogs/Bloor-on-IT-security/2010/04/mobile-phone-hacking-for-1000.html

Cloud-Based Privacy Aware Preference Aggregation Service

Sourya Joyee De and Asim K. Pal

Management Information Systems
Indian Institute of Management Calcutta, Joka, D. H. Road, Kolkata 700 104, India
sjoyeede@gmail.com, asim@iimcal.ac.in

Abstract. Each day newer security and privacy risks are emerging in the online world. Users are often wary of using online services because they are not entirely confident of the level of security the provider is offering, particularly when such services may involve monetary transactions. Often the level of security in the algorithms underlying online and cloud-based services cannot be controlled by the user but is decided by the service provider. We propose a cloud-based Privacy Aware Preference Aggregation Service (PAPAS) that enables users to match preferences with other interested users of the service to find partners for negotiation, peer-groups with similar interests etc while also allowing users the ability to decide the level of security desired from the service, especially with respect to correct output and privacy of inputs of the protocol. It also lets users express their level of trust on the provider enabling or disabling it to act as a mediating agent in the protocols. Along with this we analyze the security of a preference hiding algorithm in the literature based on the security levels we propose for the PAPAS framework and suggest an improved version of the multi-party privacy preserving preference aggregation algorithm that does not require a mediating agent.

Keywords: Security, privacy, preference aggregation, cloud computing, multi-party computation.

1 Introduction

The use of online services such as auctions, banking, shopping etc are ever-increasing. The advent of the cloud has led to the growth of these services by giving service providers access to so-called unlimited computing power, storage and the benefits of pay-per-use. However, newer security and privacy risks are also emerging each day. Users are often wary of using online services because they are not entirely confident of the level of security the provider is offering, particularly when such services may involve monetary transactions. Online services such as auctions etc hardly provide users the choice of controlling their security requirements for the underlying algorithms. Users must solely depend on the reputation of the service provider to accept that the outcome is correct and his private inputs have not been revealed to anybody else. Intense research in secure multi-party computation in the past decades

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 208–223, 2013.

has been able to provide some solutions in this respect [2, 3, 4, 6, 8, 9, 10]. However, online services still do not allow users to choose their desired level of security. It may vary from user to user and the type of service being provided. In this age of rapidly growing incidents of information security breaches, providing users the ability to specify the desirable level of security can become a critical success factor for a service provider. Moreover, different users may also have different levels of trust on service providers, in our case a CSP, based on the purpose of use of such services [5].

In this paper we propose a privacy-aware preference aggregation service based on cloud which allows users to choose the level of security they desire from the underlying preference aggregation protocols as well as specify its level of trust for the cloud service provider, hence allowing or disallowing it to act as mediating agents in the preference aggregation protocols. Preference aggregation can be an important service before two or more persons enter into a negotiation process. Before a negotiation starts, users are interested in knowing the range of persons who share a common preference with them and with whom negotiations can be entered into. They would also like to find out the common set of alternatives X in which all the negotiating parties are interested. The final interest would be to find out PO, the pareto-optimal subset of X, which will form the basis of negotiation. Similarly, persons interested in finding friend groups or peers with similar interests can use a privacy aware preference aggregation service to do so without revealing their preferences to each other before the outcome is known. The work of [4] has dealt with the problem of privacy preserving preference aggregation (although called preference hiding scheme in [4]). Our work is entirely based on the protocols they suggest. The privacy-aware preference aggregation service (PAPAS) that we propose introduces the concept of allowing users to select desired level of security from the aggregation service and uses a modified version of preference hiding algorithm of [4]. Apart from formulating the PAPAS framework, we extensively analyze the security flaws of the algorithm in [4] under the security levels we propose in this framework and then suggest our version of the preference aggregation algorithm.

Why it is necessary to give users the freedom to choose security levels of an algorithm? Would they not simply go for the highest possible security level? If a certain protocol has to be run by the user itself, on its own device, then for him efficiency of the protocol may be a great concern, especially if he is using a resource constrained device. He will therefore make a trade-off between how efficient he wants the protocol to be and how much security is required for the purpose for which he is using the protocol. If a lower level of security satisfies his requirement, then he will hardly want to run a highly secure protocol that consumes all his resources. If the protocol is run on third-party resources (such as pay-per-use clouds), then cost will be a consideration for the user. If by choosing a lower level of security, his purpose is served then he will not want to go for a highly secure protocol that costs him much more. Therefore it has been our aim to provide users with some choice about how secure he wants the protocols to be, taking into account issues such as cost and efficiency.

2 Privacy Aware Preference Aggregation Service (PAPAS)

Let us look at some more examples to convince ourselves about the need of PAPAS.

Bob is interested in buying a red Chevrolet Optra but Cathy, the car-dealer wants to sell the white variety of this car model that has been in her showroom for quite a while. Now both of them wish to know whether their preferences match and further negotiations are possible or not. But, Cathy does not want to reveal her compulsion about selling the white car because that may lead Bob to try negotiating a lower price for the car. Similarly Bob is unwilling to reveal his preference of a red car as Cathy may take advantage of his preference by demanding a higher price. Bob may even wish to search for car-dealers who have selling preferences that match his.

The movie theatre is going to play six movies A, B, C, D, E and F in the weekend. Alice, Melissa and Peter would like to watch a movie together but each of them has a different preference of movies. Alice's preference can be expressed as follows: $D > E > F > B > C > A$ where '$>$' indicates 'preferred over'. Similarly, Melissa's preference is: $E > F > D > B > C > A$ and Peter's preference is: $D > C > E > F > A > B$. However, none of them wants to reveal his/her preference to anybody else. Now, the three friends want to find out the movies that they all can watch together such that no other choice would make any of them better off without making at least one of the others worse off. This means that finding the Pareto Optimal set of movies will solve their problem. So, if movie D is preferred over movie F by two of them and not by the third, then the aggregated preference says that D and F are incomparable. This is an all-or-nothing scenario where the overall preference is 'indifferent' or 'incomparable' if different participants have different preferences over the same alternatives and the overall preference is same as the preference of the participants when all of them have the same preference over same alternatives. After aggregation of preferences of all individuals for different pairs of alternatives, a suitable comparison rule (for e.g. the comparison algorithms in [4]) can find out the desired Pareto-optimal set of movies.

The next example looks at the allocation of apartments to members of a cooperative housing society which has built a complex of apartments. The members can negotiate on prices. The negotiation process becomes simpler and faster if redundant options are removed through preference aggregation and obtaining the Pareto-optimal set.

For yet another example consider a builder who has a few alternative sites each having its pros and cons. There are several possibilities of constructions (amenities and apartment sizes, etc.) along with cost and time implications for each of these sites. The builder wants to find the customer preferences before he goes on to select a particular plan for a particular site. The problem becomes more complex as he does not have any fixed set of customers to talk to. This problem can possibly be reduced to some repeated applications of preference aggregation algorithm.

In both the above scenarios, a Privacy Aware Preference Aggregation Service (PAPAS) allows the user to come in contact with other users of the service with the same preferences as him/her without revealing his/her preferences to any other user or the cloud. This cloud-based service brings the advantage of being able to establish

contact with other users of the service who have similar preferences but were previously unknown to this user while the cloud can execute the preference aggregation algorithms either as a mediating agent or perform multi-party computations on behalf of the participants. The second example may involve running a preference aggregation algorithm several times (for aggregating preferences for pair-wise alternatives) in which case the cloud is a potential resource provider. So participants with resource constrained devices such as smart phones can easily use PAPAS whenever they wish and wherever they are.

PAPAS compares different participants' or Decision Making Agents' (DMA) preferences of one alternative a over another alternative b and finds whether all of them prefer a over b (denoted by $a > b$) or b over a (denoted by $a < b$) or some of them prefer a over b and some prefer b over a (denoted by $a \sim b$) without revealing any individual's preference to others, except what may be inferred from the output itself. We call this service 'privacy aware' because the user has the flexibility to choose the level of privacy-protection he desires from the service. A privacy aware service may protect the privacy of the user's preference from adversaries of different strengths as specified by the user. As an example, we may say that in the two illustrative scenarios above, Bob may desire a higher level of privacy of personal data than Alice as he is using the service for a purpose that may lead to a large monetary transaction as opposed to Alice's case. PAPAS enables users from different parts of the world to come in contact with each other based on their preferences in a particular context and then engage in further preference aggregation among themselves either with or without the help of the service provider as a mediating agent (MA). In this paper we have only considered 'strict' preference for the preference hiding algorithm. This may not be much meaningful if there are a large number of participants. There two possibilities, either the common set of alternatives is too small and hence of little interest, or the set is not so small, but then there will be hardly any alternative which is not in the Pareto-optimal set, which also is of little consequence. In such situations we need a relaxed constraint on the 'preference' condition, e.g. majority rule. Further as already mentioned, PAPAS can also include additional services like finding intersections, comparison schemes, participants offering alternatives (with linguistic support from the service provider for uniform framing of the alternatives, for removing redundant alternatives, etc.), allowing time zone to connect people from different places, and validating the participants (through registration and authentication).

2.1 PAPAS Framework

The PAPAS framework has three layers: 1) User Interface; 2) Interpretation Middleware and 3) Operations. The first module is a user facing module that enables the user to interact with the service by allowing them to enroll in the service, specify different requirements and inputs and obtain outputs. The Interpretation Middleware interprets all forms of user specifications to enable the right choice of protocols and participants. Thirdly, the Operations layer provides the complete set of Privacy Preserving Preference Aggregation (PPA) protocols and participant choice of which only some are triggered by the Interpretation Middleware.

User Interface. The user of a preference aggregation service is concerned about the security of the preference aggregation protocol especially the privacy of his preferences. The user may also like to choose whether the cloud should participate as a mediating agent (MA) during the execution of PPA protocol depending on how much it trusts the cloud. Apart from these, the user may also like to specify an initial quality of the other participants or DMAs based on some attributes. Users interact with PAPAS through the user interface which contains the following sub-modules: 1) Enrollment; 2) Security Requirement Statement; 3) Initial Participant Requirement Statement and 4) Preference I/O (and also inputting the alternatives with linguistic support for uniform formulation of alternatives and eliminating removing alternatives).

The options for user's specifications or requirements must be simply and comprehensively expressed so that the user has no difficulty in understanding or choosing the right one. We note here that this idea is very similar to a Service Level Agreement (SLA) but is not the same. As we have already seen, PAPAS can be used for a wide range of purposes and the security and participant requirements for each use may widely vary with the particular instance of use. The user specifies its security requirements a Security Requirement Statement (SRS). Each SRS is specific to the current use of the service and is valid till the current use is over. Similarly, the user may often wish to set a quality of participants he desires for preference aggregation. For example, Bob may want to aggregate his preferences with only a well-known car dealer in city X instead of any car dealer anywhere in the world. Therefore an initial filter for participants based on certain attributes is fixed using the Initial Participant Requirement Statement. The Enrollment module enables enrollment of new users to PAPAS whereas the Preference I/O module takes the user's preference as input and provides the user with the desired output of preference aggregation.

Interpretation Middleware. The specifications in the SRS are interpreted by a Security Requirements Interpretation Service (SRIS) to translate user security requirements to protocol security requirements. Protocol security requirements enables the Interpretation Middleware to trigger the right PPA protocol based on strength of adversarial model it uses. Similarly the Participant Requirements Interpretation Service (PRIS) translates participant requirements so that enrolled participants can be filtered and enrolled users satisfying the requirements can be allowed to participate in the particular user instance of a chosen PPA protocol. Protocol security requirements enables the Interpretation Middleware to trigger the right PPA protocol based on strength of adversarial model it uses.

Operations. This layer consists of the most important requirements for the PAPAS framework: the complete set of PPA protocols and the choice of participants as gathered from the user by Interpretation Middleware. This layer is responsible for the execution of the triggered PPA protocol with chosen participants and communicating the outcome back to the User Interface.

2.2 The User Perspective

In this section we describe the user's view of the security requirements. Therefore we first express the adversarial models from the users' viewpoint both with respect to

other participants and the CSP which may act as the mediating agent (MA) depending on user specification. Next, we specify the attributes for initial participant selection.

User-Specified Security Requirements for DMAs. We consider that DMAs do not deviate arbitrarily from the protocol i.e. behave maliciously in true sense of the term as used in the literature of SMC. However, they may perform certain actions that are not attributed to semi-honest behavior. When a DMA is semi-honest, it does not deviate from the protocol and gathers information about user inputs by keeping track of intermediate steps [1]. Sometimes DMA s may collude with others or generation of fake random numbers instead of random ones when so desired by the protocol. Moreover, an adversary can corrupt DMA s adaptively i.e. one by one as the protocol proceeds and can thus gain more information than a non-adaptive adversary. Therefore, we allow the users to base their security requirements on whether 1) DMAs collude; 2) DMA s generate fake random numbers instead of random numbers or 3) DMA s are adaptively corrupted. However, we do not require the users to understand these concepts and therefore we represent the adversarial models in terms of the strength of security a user may desire. So users are only required to choose from among the three options 1) Semi-honest DMA which indicates the lowest security level; 2) Medium Semi-honest DMA that indicates a medium security level and 3) Strong semi-honest DMA which indicates a strong security level. We shall indicate in a later section how these user security requirements are mapped to protocol security requirements by the SRIS component of the Interpretation Middleware.

User-specified Security Requirements for Cloud (MA). The user's main concern is the privacy of his preference from other users or DMAs. However since the CSP offers PAPAS with regard to preference hiding, users impose some minimal levels of trust for the cloud. Accordingly, the user may either require the CSP to act as an MA in the PPA protocol used or require the CSP to only use such PPA protocols that do not require an MA. In the later case, the Virtual Machines (VMs) in the CSP engage in a multi-party computation without any mediating VM (referred to as VM MA) while in the former case they take the help of a VM MA. When the CSP acts as an MA, the cost to users will be much less than when several VMs engage in multi-party computation. In both cases, the user tries to protect itself from DMA s of varied level of corruption but finally the user has to make a trade-off between his trust on the CSP as an MA and the costs involved in choosing the CSP to act as an MA and using VMs for multi-party computations.

Thus for the SRS, the user must specify two requirements on the CSP. First, whether it wishes the CSP to participate as an MA or not and second, if the CSP is required to participate as an MA then what level of trust the CSP is assigned i.e. 1) semi-trusted MA and 2) untrusted MA. We must mention here that the Interpretation Middleware is always assumed trusted in the sense that it will always interpret user requirements correctly and trigger the right protocol according to user requirements. It will not try to manipulate the user security requirements, for e.g., by setting the CSP to be semi-trusted MA when actually the user chose untrusted MA.

When the cloud is chosen to act as an MA, then all DMAs must provide their inputs to the MA after suitable modifications and the MA after computing the

aggregated preference will declare the output to the DMAs. On the other hand, when the cloud cannot act as an MA, each DMA is assigned a VM. The VMs, on behalf of the DMAs, exchange messages and perform the required computations. In this scenario, each VM is assumed to have the same adversarial characteristics as the DMA to which it is assigned.

User-specified Initial Participant Requirement. The user must specify what kind of participants it wants to match its preference with on the basis of different attributes. However, the attributes will depend on the nature of use of PAPAS. If the user is using PAPAS to search for car-dealers with similar preferences then the attributes required by him to filter his initial participants will be much different from the ones he requires if he is using PAPAS for movie preference matching. Therefore, the user has to first specify the broad class of participants he is looking for. During enrollment all users are provided the option to include himself in different classes of participants based on his interest in searching and being searched for preference matching. Once this broad class is specified, the user gets options suitable to the class he specified. If he specified 'car-dealers' as his interest then the attribute options he gets are location, reputation etc of car-dealers and he must specify the required level for each of these attributes.

2.3 The Protocol Perspective

The PPA protocols used in PAPAS can tolerate one of the following adversarial behaviors: 1) there is no collusion i.e. only one DMA acts as a non-adaptive adversary and it generates random numbers when required by the protocol (NC/R/NA); 2) a single DMA or a group of DMAs (i.e. collusion does not matter) is controlled by the adversary who is non-adaptive and the adversary generates numbers of its choice when actually the protocol requires random number generation (NR/NA); 3) there is collusion i.e. a group of DMAs is controlled by the adversary who is non-adaptive and the adversary generates random numbers when required by the protocol (C/R/NA); 4) there is collusion i.e. a group of DMAs is controlled by the adversary who is adaptive and the adversary generates random numbers when required by the protocol (C/R/A) and 5) a single DMAor a group of DMAs (i.e. collusion does not matter) is controlled by the adversary who is adaptive and the adversary generates numbers of its choice when actually the protocol requires random number generation (NR/A).

Table 1. Mapping Table from User Security Requirements for DMA to Protocol Security Requirements for DMA

User Security Requirements for DMAs	Protocol Security Requirements for DMAs
Semi-honest DMA	NC/R/NA
Medium Semi-honest DMA	C/R/NA, NR/NA
Strong Semi-honest DMA	C/R/A, NR/A

The SRIS component of the Interpretation Middleware uses the following mapping table to translate user security requirements for DMAs to protocol security requirements for DMAs.

The PRIS component of Interpretation Middleware finds out the subset of enrolled users with which a user can participate in a PPA protocol based on the attributes specified by the user. So it defines which DMAs can participate in a particular use by a particular user of a PAPAs service.

3 Privacy Preserving Preference Aggregation Protocols

Parties (persons, organizations etc) in a negotiation process want to reach a consensus without revealing their constraints on the acceptability and desirability of the available choices thus preventing unacceptable information disclosure about the business operations or strategies that they have adopted. If the information exchange essential to the negotiation process occurs through electronic media, then the process is called e-negotiation. Negotiators may also take help of support tools to arrive at a better decision or the whole process of negotiation may be fully automated [7].

[4] attacks the problem of finding the Pareto-optimal frontier in multi-party negotiations allowing minimum information disclosure using the solutions to secure multi-party computation problems in set theory, linear programming etc. Therefore, various privacy preserving algorithms for the different steps of solving this problem have been discussed. The steps are: 1) finding out the feasible space by using intersection algorithms and 2) finding the Pareto-optimal subset. The algorithms for the latter step consists of two major components 1) Preference Hiding Schemes and 2) Comparison Schemes. After the Comparison Scheme (that repeatedly calls the Preference Hiding Protocol), a set of non-dominated alternatives i.e. the Pareto-optimal set is arrived at. The algorithms have been designed considering that the negotiating parties i.e. the decision making agents (DMAs) and the mediating parties i.e. the mediating agents (MA) in the negotiation process are semi-honest. There is no extensive discussion on the effect of collusions etc on the algorithms or different possible security scenarios. We take this opportunity to use the preference hiding scheme proposed by [4] and propose a Cloud-based Privacy Aware Preference Aggregation Service by taking into account the several possible adversary models the user may like to consider while using such a cloud-based service.

3.1 XOR-Based Preference Aggregation Protocol

The XOR-based Preference hiding Algorithm proposed by [4] is not secure under the adversarial models we have considered as there is either partial or full information disclosure (See Appendix for details). We present here a security analysis of their protocol based on the adversary models we have proposed and suggest a modification of this algorithm that will remove the drawbacks. The main motivation behind the modification is to remove the role of the central DMA which proves to be a notorious give-away.

Before we analyze the security of the algorithm considering the adversarial models as proposed in section 2.3, we provide a brief overview of the algorithm. One of the participants i.e. DMAs (call it the central DMA or DMA^c) generates random representations e_1^i for the preference relation $a > b$ and e_2^i for the preference relation $a < b$ for each DMA^i $(i = 1 \ldots \ldots m)$ distributes these representations to the corresponding DMA. Next each DMA^i generates $m - 1$ random numbers and by using these random numbers as masks, computes $h_+ = \oplus_i e_1^i$ to indicate $a > b$ and $h_- = \oplus_i e_2^i$to indicate $a < b$ in a distributed manner. Afterwards, depending on its preference, DMA^i sets the value of its preference e^i to be either e_1^i or e_2^i. The earlier step using random numbers as mask is repeated to calculate in a distributed manner the aggregated preference i.e. $h = \oplus_i e^i$. We shall see that information disclosure occurs mainly because of the fact that the single DMA, the central DMA or DMA^c, generates the random pairs of binary vectors that correspond to the choices of each DMA. This makes DMA^c powerful in terms of the information about how each DMA's choice is represented. Notably, the other DMA s only know the representation of their own choices and nothing about how other DMA s' choices are represented. For DMA^c, deducing some information about the choice of at least some of the other DMAs, is not very difficult as we will show below. Also, when DMA^c retains the right to generate the random representations of the preferences, it may as well calculate the values of h_+ and h_- itself and distribute it to the other DMA s. The only advantage of each DMA calculating it on its own is that DMA^c would not be able to manipulate this calculation somehow or give incorrect or different values for h_+ and h_- to the DMA s. The authors in [4] do not clearly mention the adversary model they have considered.

We must note here that DMA^c should either be elected or randomly selected at the beginning of each instance of the algorithm (although authors in [4] do not mention this). In the whole negotiation process, the comparison of alternatives occurs a number of times and each time the preference hiding algorithm is to be used to secretly compute the overall preference of the DMA s. We may have to either assume that an adversary controlling a fixed number of DMA s exists before the negotiation process begins (non-adaptive) or that the adversary can select whom to corrupt as the protocol proceeds. The adversary's success lies in its ability to either begin with such a set of parties one of whom will be selected as DMA^c or corrupt the party chosen as DMA^c later on. If a DMA is randomly selected to be DMA^c, then beginning with a fixed set of corrupted parties does not guarantee that will belong to that set. If a DMA is elected to the position of DMA^c, then the adversary may influence the election process in such a way so that one of the colluding DMA s gets selected as DMA^c. In case of adaptive adversary, however, corrupting the central DMA later is easier than in the non-adaptive case.

For the rest of the analysis, we consider that there are a total of m DMA s of which c collude and b bits are used to represent the choices of each DMA. For results, we shall see that whenever we consider that adversaries can generate fake random numbers without any other deviation from the protocol, full information disclosure occurs. Also since the success of the adversary depends on whether it is able to corrupt DMA^c, the adaptive adversary gains a clear advantage over non-adaptive

ones. Collusion helps the adversary to some extent when it is absolutely non-deviating (i.e. C/R/NA). Below we present the analysis for each adversary model.

NC/R/NA. In this case, DMA^c will generate the binary vectors randomly but will still be able to know some information about the choices of some of the other DMA s. We show this by an example (see Appendix). However, DMA^c will not know beforehand for which of the DMAs it will be able to know the exact choice. This is a disadvantage for it. However, if DMA^c is chosen randomly from the set of DMAs, the chances of the adversary of corrupting DMA^c becomes very low. The adversary gains no information by corrupting any other DMA.In the worst case, full information disclosure is possible (for explanation, see Appendix).

NR/NA. In this scenario, the adversary (if it is DMA^c) instead of generating random representations of preferences, can construct them in such a way so as to help it in maximizing its information gain. We see that in this scenario causing full information disclosure is very easy and it is independent of the presence of collusion. In this case also, such disclosure is possible only if the adversary is able to corrupt DMA^c chances of which are very low for reasons we have already mentioned.

C/R/NA. The collusion benefits only when DMA^c is a part of it. A collusion of c DMAs including DMA^c makes it easier for the adversary to derive information about the exact choice of at least some of the non-colluding DMAs. So we see here that with collusion that includes DMA^c, there is quite a significant information disclosure. In the worst case, full information disclosure is possible.

C/R/A. This case is similar to C/R/NA with the exception that the adversary is now able to choose whom to corrupt at any point during the protocol execution which implies that it is able to corrupt DMA^c almost certainly once the latter has been selected randomly. Information disclosure is same as that of C/R/NA.

NR/A. This case is similar to NR/NA with the exception that the adversary is now able to choose whom to corrupt at any point during the protocol execution which implies that it is able to corrupt DMA^c almost certainly once the latter has been selected randomly. This situation is definitely the worst possible scenario as it leads to full information disclosure with very high probability. The information disclosure is independent of whether there is collusion or not.

Now we propose our version of the XOR-based Preference Aggregation Protocol that we use as one of the PPA protocols in PAPAS. We use the same notations as in [4].

Our XOR-Based Preference Aggregation Protocol without MA

Input: Alternatives a and b
Output: $a > b$ or $a < b$ or $a \sim b$
Condition: No mediators

Set up Phase: Each DMA^i generates a pair of b-bit binary vectors $(e_1^i, e_2^i), i = 1, \ldots\ldots, m$. The number of bits denoted by b is fixed previously depending on the number of DMAs and is sufficiently large. This vector pair is to be kept secret by each DMA.

Calculation of h_+ and h_-:

The DMAs need to calculate $h_+ = \oplus_i e_1^i$ to indicate $a > b$ and $h_- = \oplus_i e_2^i$ to indicate $< b$. This can be done as follows:

Each DMA^i

1. Generates $m-1$ random vectors $r_1^i, \ldots\ldots\ldots, r_{m-1}^i$.
2. Finds $r^i = e_1^i \oplus r_1^i \oplus r_2^i \oplus \ldots\ldots\ldots \oplus r_{m-1}^i$.
3. Sends randomly one vector p^{ij} from the set of vectors $\{r_1^i, \ldots\ldots\ldots, r_{m-1}^i, r^i\}$ to each DMA^j where $j = 1, \ldots\ldots, m; j \neq i$ and retains one with itself. No vector is sent to more than one DMA.
4. Finds XOR of the vectors received in the previous step i.e. it obtains $\alpha^i = \oplus_j p^{ji}$ and sends it to all other DMAs.
5. Finds XOR of the vectors received in the previous step i.e. it obtains $h_+ = \oplus_i \alpha^i$.

The above process is repeated for calculation of h_-.

Preference Aggregation Phase:

Each DMA^i performs the following:

1. Finds $e^i = \begin{cases} e_1^i \text{ when } v^i(a) \geq v^i(b) \\ e_2^i \text{ when } v^i(a) < v^i(b) \end{cases}$ where (e_1^i, e_2^i) is a random pair of binary vectors received in the Set up phase.
2. Generates $m-1$ random vectors $r_1^i, \ldots\ldots\ldots, r_{m-1}^i$.
3. Finds $r^i = e^i \oplus r_1^i \oplus r_2^i \oplus \ldots\ldots\ldots \oplus r_{m-1}^i$.
4. Sends randomly one vector p^{ij} from the set of vectors $\{r_1^i, \ldots\ldots\ldots, r_{m-1}^i, r^i\}$ to each DMA^j where $j = 1, \ldots\ldots, m; j \neq i$ and retains one with itself. No vector is sent to more than one DMA.
5. Finds XOR of the vectors received in the previous step i.e. it obtains $\alpha^i = \oplus_j p^{ji}$ and sends it to all other DMAs.
6. Finds XOR of the vectors received in the previous step i.e. it obtains $h = \oplus_i \alpha^i$.
7. If $h = h_+$ then $a > b$ else if $h = h_-$ then $a < b$ else $a \sim b$.

When more comparisons are to be made then this whole process is to be repeated.

Security Analysis. The adversary can take advantage of the XOR-based algorithm by [4] because of the use of DMA^c which assumes the important role of generating the representation of the choices of each of the DMAs. The role of DMA^c has been removed in our algorithm and each DMA is now allowed to randomly generate its own binary vector i.e. the representation of its own choice. This ensures that none of the

DMAs will have with it all the representations and hence knowing the individual choices becomes impossible even when some of them collude. No information disclosure takes place under adaptive or non-adaptive adversaries.

4 Conclusion and Future Work

In this paper we have presented a Cloud-based Privacy Aware Preference Aggregation Service that allows enrolled users to search for other unknown users of the service and find out whether preferences in a certain context match. The preference matching can occur among multiple users depending upon the use-case and the user has the flexibility to choose the security level of the protocols used to prevent privacy-breach of his preferences and the types of participants preferred for the preference aggregation. Our framework may be looked upon as a precursor to a cloud based Privacy Aware Negotiation-as-a-Service that will allow users who previously did not know each other to negotiate on their preferences to reach a consensus without revealing their preferences to anyone. Algorithms of different security level, efficiency and with different preference aggregation rules can be integrated into this service. We are working towards this end.

Existing online auction services such as eBay.com etc do provide a certain platform of negotiation between buyers and sellers but they focus on a very specific use i.e. buying and selling of goods and do not allow the flexibility to users to choose the security level of the service. In contrast, PAPAS is a highly flexible cloud-based service which allows users to choose their desired level of security and types of participants they want to interact with apart from providing a generalized arena for preference aggregation on any use-case starting from choice of movies, restaurants among friends to buying and selling of cars, real estate etc. The process of choosing security levels for algorithms must be made comprehensive for users so that users with little or no knowledge about security can effectively express their choices.

References

1. Goldreich, O.: Foundations of Cryptography Volume II Basic Applications. Cambridge University, Cambridge (2004)
2. Chakraborty, S., Sehgal, S.K., Pal, A.K.: Privacy preserving e-negotiation Protocols based on secure multiparty computation. In: Proceedings of the IEEE SoutheastCon 2005, pp. 455–461. Springer (2005)
3. Saroop, A., Sehgal, S.K., Ravikumar, K.: Multi-Attribute Auction Format for Procurement with Limited Disclosure of Buyer's Preference Structure. In: Decision Support for Global Enterprises Annals of Information Systems, vol. 2, pp. 257–267. Springer (2007)
4. Sehgal, S.K., Pal, A.K.: Finding Pareto Optimal Set of Distributed Vectors with Minimum Disclosure. In: Sen, A., Das, N., Das, S.K., Sinha, B.P. (eds.) IWDC 2004. LNCS, vol. 3326, pp. 144–149. Springer, Heidelberg (2004)

5. De, S., Saha, S., Pal, A.K.: Achieving Energy Efficiency and Security in Mobile Cloud Computing. In: Proceedings of the 3rd International Conference on Cloud Computing and Services Sciences CLOSER 2013. SciTePress (2013)
6. Du, W., Zhan, Z.: A Practical Approach to Solve Secure Multi-party Computation Problems. In: Proceedings of the 2002 workshop on New security paradigm, NSPW 2002, pp. 127–135. ACM (2002)
7. Bichler, M., Kersten, G., Strecker, S.: Towards a Structured Design of Electronic Negotiations. In: Group Decision and Negotiation, vol. 12(4), pp. 311–335 (2003)
8. Naor, M., Pinkas, B., Sumner, R.: Privacy Preserving Auctions and Mechanism Design. In: Proceedings of the 1st ACM Conference on Electronic Commerce, pp. 129–139. ACM (1999)
9. Cramer, R., Gennaro, R., Schoenmakers, B.: A secure and Optimally Efficient Multi-Authority Election Scheme. European Transactions on Telecommunications 8(5), 481–490 (1997)
10. Kamara, S., Raykova, M.: Secure Outsourced Computation in Multi-tenant Cloud. In: IBM Workshop on Cryptography and Security in Clouds, WCSC 2011 (2011)

Appendix A

Here we present in details the security analysis of XOR-based preference hiding scheme of [4] under NC/R/NA and NR/NA we propose. We have not shown the rest for lack of space.

NC/R/NA. Let us suppose that there are ten DMAs among which one has been elected or randomly chosen as DMA^c. Now, DMA^c (say DMA^5 is DMA^c) generates the binary vector pair $(e_1^i, e_2^i), i = 1, \ldots\ldots, m$ randomly. We assume that each vector is of 10-bit length.

After DMA^c randomly distributes these vector pairs to each of the DMAs, the following is the random allocation:

Table 2. Randomly Generated Vectors for Input Representation as distributed to each DMA by DMA^c

DMA	e_1^i	e_2^i
DMA^1	00 1001 1011	01 0101 0101
DMA^2	01 1110 1001	11 1010 1111
DMA^3	00 0011 0101	10 0110 1001
DMA^4	11 1111 1000	10 1010 0000
DMA^5 (DMA^c)	11 0110 0101	00 0010 1100
DMA^6	10 1100 1010	01 1111 1010
DMA^7	10 0000 0011	00 0101 0011
DMA^8	10 0101 1100	01 1100 0000
DMA^9	01 1111 1001	11 0111 0001
DMA^{10}	00 1000 0001	00 1001 0111

h_+ and h_- are calculated as follows:

$$h_+ = \oplus_i e_1^i = 10\ 0011\ 0111$$
$$h_- = \oplus_i e_2^i = 01\ 1001\ 0000$$

At the end of the algorithm, each of the DMAs comes to know about $h = \oplus_i \alpha^i = \oplus_i e^i$

Let us now tabulate the choices of the DMAs.

Table 3. Choices of each DMA

DMA	Choices
DMA^1	00 1001 1011
DMA^2	01 1110 1001
DMA^3	10 0110 1001
DMA^4	11 1111 1000
DMA^5 (DMA^c)	11 0110 0101
DMA^6	01 1111 1010
DMA^7	00 0101 0011
DMA^8	01 1100 0000
DMA^9	11 0111 0001
DMA^{10}	00 1000 0001

Given these choices, we have $h = \oplus_i \alpha^i = \oplus_i e^i = 00\ 0001\ 1111$.

Now given the values of h and $\left(e_1^i, e_2^i\right)$ we show below that it is possible for DMA^c to find out the exact choices for at least some of the DMAs. The following table shows the bits (represented by X) where the representations of each of the choices of each DMA (except DMA^c) vary. Also we may consider that this table is available to DMA^c.

Table 4. Adversary's Analysis Table

DMA	Varying bits in Choice
DMA^1	0X XX01 XXX1
DMA^2	X1 1X10 1XX1
DMA^3	X0 0X1X XX01
DMA^4	1X 1X1X X000
DMA^5 (DMA^c)	11 0110 0101
DMA^6	XX 11XX 1010
DMA^7	X0 0X0X 0011
DMA^8	XX X10X XX00
DMA^9	X1 X111 X001
DMA^{10}	00 100X 0XX1
Value of h	**00 0001 1111**

Using the property of XOR that odd number of true values (1s) when XORed gives true value (1), DMA^c can deduce the following from the table above: 1) the second bit of one or all of DMA^1, DMA^2 and DMA^{10} are 1; 2) the third bit of DMA^1, DMA^2,DMA^3, DMA^8 and DMA^{10} are all 0 or that of two or four of them are 1; 3) the fourth bit of three or all of DMA^1, DMA^3, DMA^4, DMA^8 and DMA^9 are 1 etc.

Although DMA^c comes to know these pieces of information, it is difficult for it to know the exact choice of any of the other DMAs. Since the above table can be derived only by DMA^c (because only DMA^c knows all the options of all the DMAs), no other DMA can derive these pieces of information.

The worst case here will arise when the adversary generates the representations of choices such that within each pair, the choices differ by a single bit and across pairs the position of difference varies. However, since the pairs have to be randomly generated, the probability of being able to generate the representations according to the above condition will be $\frac{b.(b-1).....(b-m-1)}{b^m}$ where m is the number of DMAs and b is the number of bits used in the representations of choices. We must note here that $2^b >> 2m$.

NR/NA. In this scenario instead of generating $\left(e_1^i, e_2^i\right), i = 1, \ldots\ldots, m$ randomly, the adversary (if it is DMA^c) can construct the vectors in such a way so as to help it in maximize its information disclosure. We give an example below.

Let us consider that the adversary constructs and distributes the binary vectors according to the following table:

Table 5. Vectors for Input Representation

DMA	e_1^i	e_2^i
DMA^1	01 0000 0000	11 0000 0000
DMA^2	00 1000 0000	01 1000 0000
DMA^3	00 0100 0000	00 1100 0000
DMA^4	00 0010 0000	00 0110 0000
DMA^5 (DMA^c)	00 0000 0000	11 1111 1111
DMA^6	00 0001 0000	00 0011 0000
DMA^7	00 0000 1000	00 0001 1000
DMA^8	00 0000 0100	00 0000 1100
DMA^9	00 0000 0010	00 0000 0110
DMA^{10}	00 0000 0001	00 0000 0011

$$h_+ = \oplus_i e_1^i = 01\ 1111\ 1111$$
$$h_- = \oplus_i e_2^i = 01\ 1111\ 1110$$

Let the actual choices of the DMAs be as follows:

Table 6. Choice of each DMA

DMA	Choice
DMA^1	01 0000 0000
DMA^2	01 1000 0000
DMA^3	00 1100 0000
DMA^4	00 0010 0000
DMA^5 (DMA^c)	00 0000 0000
DMA^6	00 0011 0000
DMA^7	00 0001 1000
DMA^8	00 0000 1100
DMA^9	00 0000 0110
DMA^{10}	00 0000 0001
Value of h	**00 0100 0011**

Given these choices, we have $h = \oplus_i \alpha^i = \oplus_i e^i = 00\ 0100\ 0011$.

Now, as before, the adversary deduces the following table from its knowledge about the representation of the choices of each of the DMAs.

Table 7. Adversary's Analysis Table

DMA	Choice
DMA^1	X1 0000 0000
DMA^2	0X 1000 0000
DMA^3	00 X100 0000
DMA^4	00 0X10 0000
DMA^5 (DMA^c)	**00 00X1 0000**
DMA^6	00 0000 0000
DMA^7	00 000X 1000
DMA^8	00 0000 X100
DMA^9	00 0000 0X10
DMA^{10}	00 0000 00X1
Value of h	**00 0100 0011**

Now, it becomes easy for the adversary to deduce the exact choice made by each of the other DMAs using the property of XOR mentioned previously. So the adversary comes to know 1) the eighth bit of DMA^1 is 0 and its choice is e_1^1. Hence DMA^1 prefers$a > b$; 2) The seventh bit of DMA^2 must be 1 and its choice is e_2^2. So it prefers $a < b$. Similarly, the adversary comes to know the choices of each of the DMAs.

A Method for Re-using Existing ITIL Processes for Creating an ISO 27001 ISMS Process Applied to a High Availability Video Conferencing Cloud Scenario

Kristian Beckers[1], Stefan Hofbauer[2],
Gerald Quirchmayr[3,4], and Christopher C. Wills[5]

[1] University of Duisburg-Essen, paluno - The Ruhr Institute for Software Technology
Kristian.Beckers@paluno.uni-due.de
[2] Amadeus Data Processing GmbH, Network Integration Services
Stefan.Hofbauer@amadeus.com
[3] University of Vienna, Multimedia Information Systems Research Group
gerald.quirchmayr@univie.ac.at
[4] University of South Australia, School of Computer and Information Security
gerald.quirchmayr@unisa.edu.au
[5] CARIS Research Ltd.
ccwills@carisresearch.co.uk

Abstract. Many companies have already adopted their business processes to be in accordance with defined and organized standards. Two standards that are sought after by companies are IT Infrastructure Library (ITIL) and ISO 27001. Often companies start certifying their business processes with ITIL and continue with ISO 27001. For small and medium-sized businesses, it is difficult to prepare and maintain the ISO 27001 certification. The IT departments of these companies often do not have the time to fully observere standards as part of their daily routine. ITIL and ISO 27001 perfectly fit into companies and help reduce errors through the standardization and comparability of products and services between themselves and other companies and partners. ISO 27001 specifically looks at security risks, countermeasures and remedial actions.

We start with the processes that need to be in place for implementing ITIL in an organisation's business processes. We use a cloud service provider as a running example and compare ITIL processes with ISO 27001 processes. We identify which aspects of these two standards can be better executed. We propose a mapping between ITIL and ISO 27001 that makes them easier to understand and assists with the certification process. We show further how to prepare for audits as well as re-certification. Often, these two processes are seen separately and not in conjunction, where synergies can be exploited. Legal requirements, compliance and data security play an integral part in this process. In essence, we present checklists and guidelines for companies who want to prepare for standardization or that are already certified, but want to improve their business processes. We illustrate our method using an high availability video conferencing cloud example.

Keywords: ITIL, ISO 27001, cloud computing, processes, standards, certification, compliance.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 224–239, 2013.

1 Introduction

Before an organisation can consider becoming certified and begin applying certified standards, there must be a clarity of understanding in relation to the organisation's business processes. These processes must be defined, agreed and adopted. The organisation must understand, the service level agreements (SLA) operated by their clients or customers and which services the organisation offers as managed service provider. For audit reasons this knowledge must be reviewed twice a year and any emergent changes need to be documented.

Another important step is the introduction of a ticketing system, like open source product OTRS, to capture all customer interactions and support the processes. The ticketing system should also contain an asset inventory, where all hardware and software is itemised. Every update, addition, change or deletion must be reflected in this inventory. Additionally hardware and defined services should be monitored, perhaps with an open source software such as Nagios[1] for example.

The move towards cloud services is accelerating the rate at which companies develop and the use of such service requires the sharing of large amounts of data in a secure and scalable way. ITIL and ISO 27001 help companies to achieve customer friendliness and high quality services and support. Without standardization, a lot of manual effort, growing cost and complexity for an organization arises, as the number of their service partners grows.

An organisation should definitely start with certifying ITIL and proceed with the ISO 27001 aiming at security afterwards. The basis for the ITIL certification is an understanding of ITIL itself, quality assurance and knowledge of the business processes in place. Capacity management assures that there is no capacity shortage on the provider side, as well as on the customer side. Regular checks and warnings, based on defined thresholds help to meet the capacity requirements and plan for expected growth. Exceptionally fast growth or the adding of resources must be planned, organized and scheduled, with the cloud provider service manager. This is an important step towards supporting the client or customer. At the base of the organisational hierarchy is the first level desk, then the technical engineers, the service manager and a department manager or even chief executive officer. The involvement of these roles depends on the defined SLAs and the target-, reaction time to solve incidents.

The rest of the paper is organized as follows. Section 2 presents background on the standards ITIL and ISO 27001. Section 3 serves as structured method for Section 4, which shows how these two standards can be mapped and mutual benefit created from the synergy between each other. Section 5 shows a usage example of the method based upon a real-life example and Sect. 6 presents related work. Section 7 concludes and gives directions for future research.

2 Background

We illustrate the general idea of cloud computing in Sect.2.1, and our cloud system analysis pattern in Sect. 2.2. We introduce the ISO 27001 standard in Sect. 2.3 and the ITIL Standard in Sect. 2.4.

[1] `http://www.nagios.org`

2.1 Cloud Computing

The term *cloud computing* describes a technology as well as a business model [1]. According to the *National Institute of Standards and Technology (NIST)* cloud computing systems can be defined by the following properties [2]: the cloud customer can acquire resources of the cloud provider over *broad network access* and *on-demand* and pays only for the used capabilities. Resources, i.e., storage, processing, memory, network bandwidth, and virtual machines, are combined into a so-called *pool*. Thus, the resources can be virtually and dynamically assigned and reassigned to adjust the customers' variable load and to optimize the resource utilization for the provider.

The virtualization causes a location independence: the customers generally have no control or knowledge of the exact location of the provided resources. Another benefit is that the resources can be quickly scaled up and down for customers and appear to be unlimited, (this is called *rapid elasticity*). The pay-per-use model includes guarantees such as availability or security for resources via customized *Service Level Agreements (SLA)* [3].

The architecture of a cloud computing system consists of different service layers and allows different business models: on the layer closest to the physical resources, the *Infrastructure as a Service (IaaS)* provides pure resources, for example virtual machines, where customers can deploy arbitrary software including an operating system. Data storage interfaces provide the ability to access distributed databases on remote locations in the cloud. On the *Platform as a Service (PaaS)* layer, customers use an API to deploy their own applications using programming languages and tools supported by the provider. On the *Software as a Service (SaaS)* layer, customers use applications offered by the cloud provider that are running on the cloud infrastructure. Furthermore, cloud providers require a layer that monitors their customers' resource usage, e.g. for billing purposes and service assurances. Buyya et al. [4] introduce this layer as a middleware in their cloud model. Cloud computing offers different *deployment scenarios*: *private clouds* are operated solely for an organization, *public clouds* are made available to the general public or a large industry group and are owned by a third party selling cloud services. In between these scenarios are *hybrid clouds* where users complement internal IT resources upon demand with resources from an external vendor [1].

2.2 Cloud System Analysis Pattern

We propose patterns for a structured domain knowledge elicitation. Depending on the kind of domain knowledge that we have to elicit for a particular software engineering process, we always have certain elements that require consideration. For this work we use a specific *context elicitation pattern*, the so-called *cloud system analysis pattern* [5]. We base our approach on Jackson's work on Problem Frames [6] that considers requirements engineering from the point of view of a machine in its environment. The machine is the software to be built and requirements are the effect the machine is supposed to have on the environment. Any given environment considers certain elements, e.g., stakeholders or technical elements. Jackson [6], who describes Problem Frames as follows: *"A problem frame is a kind of pattern. It defines an intuitively identifiable problem class in terms of its context and the characteristics of its domains, interfaces*

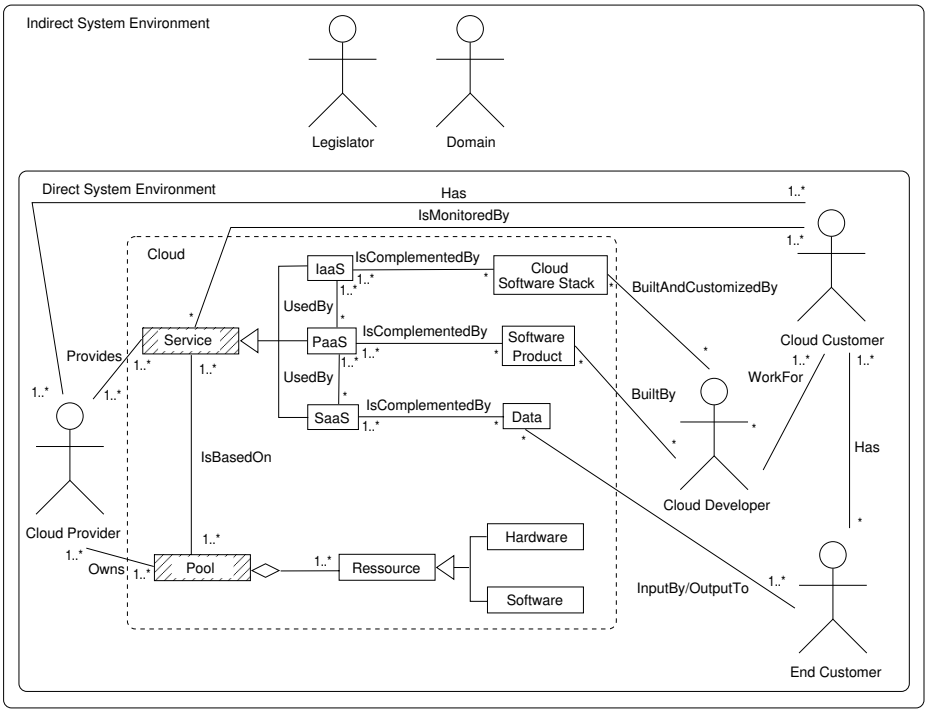


Fig. 1. Cloud System Analysis Pattern taken from [5]

and requirement.". We were also informed by Fowler [7]. Fowler, developed patterns for the analysis phase of a given software engineering process. His patterns describe organizational structures and processes, e.g., accounting, planning, and trading.

Our patterns for the analysis phase differ from patterns concerning solutions for the design phase of software engineering like the Gang of Four patterns [8] or the security patterns by Schumacher et al. [9]. The reason is that we provide a means for a structured elicitation of domain knowledge for cloud computing systems. We do not provide solutions for the implementation phase of clouds. We present a short introduction of our so-called *Cloud System Analysis Pattern (or short: Cloud Pattern)* [5] in the following. We created the pattern for cloud-specific context establishment and asset identification compliant to the ISO 27000 series of standards. A *Cloud* (see Fig. 1) is embedded into an environment consisting of two parts, namely the *Direct System Environment* and the *Indirect System Environment*. The *Direct System Environment* contains stakeholders and other systems that directly interact with the *Cloud*, i.e. they are connected to the cloud by associations. Moreover, associations between stakeholders in the *Direct* and *Indirect System Environment* exist, but not between stakeholders in the *Indirect System Environment* and the *Cloud*. Typically, the *Indirect System Environment* is a significant source for compliance requirements. The *Cloud Provider* owns a *Pool* consisting of *Resources*, which are divided into *Hardware* and *Software* resources. The provider offers its resources as *Services*, i.e. *IaaS*, *PaaS*, or *SaaS*. The boxes *Pool* and

Service in Fig. 1 are hatched, because it is not necessary to instantiate them. Instead, the specialized cloud services such as *IaaS*, *PaaS*, and *SaaS* and specialized *Resources* are instantiated. The *Cloud Developer* represents a software developer assigned by the *Cloud Customer*. The developer prepares and maintains an *IaaS* or *PaaS* offer. The *IaaS* offer is a virtualized hardware, in some cases it is equipped with a basic operating system. The *Cloud Developer* deploys a set of software named *Cloud Software Stack* (e.g. web servers, applications, databases) into the *IaaS* in order to offer the functionality required to build a *PaaS*. In our pattern *PaaS* consists of an *IaaS*, a *Cloud Software Stack* and a *cloud programming interface (CPI)*, which we subsume as *Software Product*. The *Cloud Customer* hires a *Cloud Developer* to prepare and create *SaaS* offers based on the CPI, finally used by the *End Customers*. *SaaS* processes and stores *Data* input and output from the *End Customers*. The *Cloud Provider*, *Cloud Customer*, *Cloud Developer*, and *End Customer* are part of the *Direct System Environment*. Hence, we categorize them as *direct stakeholders*. The *Legislator* and the *Domain* (and possibly other stakeholders) are part of the *Indirect System Environment*. Therefore, we categorize them as *indirect stakeholders*. We also provide templates for each stakeholder that describe their attributes like motivation for using the cloud.

2.3 The ISO 27001 Standard

The ISO 27001 defines the requirements for establishing and maintaining an Information Security Management System (ISMS) [10]. In particular, the standard describes the process of creating a model of the entire business risks of a given organization and specific requirements for the implementation of security controls. The ISO 27001 standard is structured according to the "Plan-Do-Check-Act" (PDCA) model, the so-called *ISO 27001 process* [10]. In the *Plan* phase an ISMS is established, in the *Do* phase the ISMS is implemented and operated, in the *Check* phase the ISMS is monitored and reviewed, and in the *Act* phase the ISMS is maintained and improved. In the *Plan* phase, the *scope and boundaries* of the ISMS, its *interested parties*, *environment*, *assets*, and all the *technology* involved are defined. In this phase also the ISMS *policies*, *risk assessments*, *evaluations*, and *controls* are defined. Controls in the ISO 27001 are measures to *modify risk*. The ISO 27001 standard demands the creation of a set of documents and the certification of an ISO 27001 compliant ISMS is based upon these documents.

Changes in the organisation or technology also have to comply with the documented ISMS requirements. Furthermore, the standard demands periodic audits towards the effectiveness of an ISMS. These audits are also conducted using documented ISMS requirements. In addition, the ISO 27001 standard demands that management decisions, providing support for establishing and maintaining an ISMS, are documented as well. This support has to be documented via management decisions. This has to be proven as part of a detailed documentation of how each decision was reached and how many resources are committed to implement this decision.

2.4 The ITIL Standard

The IT Infrastructure Library (ITIL) [11] is a collection of best practices for implementing an IT service management. The standard provides example processes for typical

tasks regarding IT management. The standard also provides tools of how to consider planning, establishing, supporting and optimizing of IT services in order to achieve business goals.

ITIL is a defacto standard for the creation, establishment and management of critical processes. ITIL contains generic descriptions and is independent of vendors or technology. ITIL provides a set of process that contain: basic requirements for the process, goals of the process, pattern for procedures and roles, interfaces for different processes, hints for critical success factors, suggestions for measuring key performance indicators, knowledge about success criteria for deploying the process.

3 Method

Cost, flexibility and ease of operation are driving more and more organisations into the cloud. We differentiate here between the public cloud, with cloud services of providers such as Amazon or Salesforce and the private cloud of virtualised services running on an organisation's privately run hardware and software.

It is not easy for a huge organisation to transition everything over to the cloud at once, because then parts of the business critical systems and services are already transitioned into the cloud and some are not. Cloud integration platform help solve this issue and help unite those services.

With the cloud integration platform it becomes much easier for global cloud players to sell cloud services to their partners. We consider clouds in our method via using our cloud system analysis pattern introduced in Sect. 2.2.

3.1 ITIL

Our practical experience has been, gained in different organisations, both mid-sized to large. From this experience, we identify the following steps relevant for implementing ITIL within a company. Different roles must be introduced and are needed to comply with this standard.

Define Initial Tasks for the Roles. Our method requires a service manager, who is having regular service meetings with the service provider, as well as a service desk including a service manager. The service desk covers agreed service times and may also be on duty after business hours for special reasons, such as emergency incidents. A change advisory board is established, whose task is to supervise proposed changes and confirm or cancel those changes. Changes can be cancelled for instance if no fallback is available, or the risks are too high. Field engineers work closely together with external customers to help fulfil their business needs. They are part of the service desk, but most of the time, are not at the office. The service manager, together with the service desk manager is also responsible that the customer gets a SLA report at least once a month. If the SLA criteria are not met, the service provider must reimburse as defined in the service contract between these two parties.

Create an Incident Response System. The service desk uses a ticket system, such as OTRS for incident management, as well as capacity management and inventory management. The ticketing system also keeps track of the guaranteed reaction time and time to recover for incidents. Service engineers can use the ticketing system to record their invested support time, which can then be used for billing purposes, as well as a means to communicate with the customer through e-mails. The agreed SLA values are part of the service contract and can be further negotiated by both sides, assuming the mutual consent.

Establish a Monitoring System. It is the duty of the service provider to monitor their own core infrastructure as well as customer devices. If requested, the service provider can also monitor devices within the customer site. The monitoring solution automatically generates tickets, based on the occurrence of incidents. The certified service provider needs a thorough documentation of their main business processes and service processes, including revisions on a regular basis. Every change in a productive environment, whether within the service provider context or the customer side, must be documented in a change document. Escalation processes need to be in place for incidents or problems. It must be clear, when a ticket is escalated and to whom.

Getting Certified in the Core Business Areas. It is obviously a good idea to have trained and certified personal, because it will reduce the chance of errors as well as having a better relationship to third level partners and a better reputation against customers. For the non-daily business, a project manager is supervising the project's tasks and can acquire additional resources or take other necessary decisions. To have a clear vision of the provided services the service provider should have a service catalog for the provided service, with responsibilities and boundaries to partners. This information must be agreed on and communicated, prior having a contract with a customer. Not only the software or hardware producer is acting as third level support, but also external consultants or other companies, partners for third level support. Additional feature requests in upcoming software releases can be wished for at software partners.

Document All Necessary Information. The provided service needs to be documented for technical as well as organizational aspects and revised, whenever there is a change. This documentation should be stored on a document management system, like Sharepoint, where it is easy to check out documents, make changes and check in the new document. For every customer, there needs to be a technical as well as sales person. The most important customer satisfaction parameters are availability, costs, service, innovation, upselling products as well as security. The security is of high importance for the provider's as well as customer's equipment. This includes the regular patching of devices and yearly external security audits.

Define Boundaries to Stakeholders. The customer further needs a system specification and operating procedures handbook. For customer projects a project plan, time plan and dedicated project manager is needed from service provider side. The service provider should also think about a certified quality assurance employee, who takes care after quality management and quality assurance. This person needs to communicate a lot with externals as well as all involved internal departments. If the quality of the service is not assured, it is most likely that the customer will

not continue the contract or as a worst case, cancel the contract with the service provider immediately. A service provider should have a service catalog, where all provided services and the engineers' responsibilities are listed.

3.2 ISO 27001

Once an organisation has been successfully certified in the area of ITIL, it is time to seek compliance with ISO 27001. As with ITIL, the introduction of the standard within the oranisation is made up of different sub steps.

Agree on an IT Security Policy. In order to become certified in the field of ISO 27001 an organisation has to adopt the following steps. The first requirement is that of a company security policy, which handles topics, such as choosing a secure password, rights of usage of the Internet and corporate e-mail, locking of the computer, data leakage and prevention, social engineering, storage of data and confidence regarding data and information. Regular awareness training for users and administrators is provided by instructors, in order to familiarize key personal with the security policy. After the security policy training, the attendees should give their written consent, in order that the security policy can be instituted.

Establish the Position of a CISO. Furthermore, the position of a Chief Information Security Officer (CISO) is required. This position can be led by a person, who is also responsible for other topics within the company. The tasks of a CISO includes policies, standards, procedures, architecture and guidelines for the organisation's business. The first priority is the establishment of an information security function to ensure that all the organisation's information assets are well protected and mitigations are adequately implemented. The CISO will also manage the ongoing execution of the security operation in all of the organisation's information technology areas such as applications, data protection, data communications systems as well as all information systems.

Other tasks carried out by the CISO drive the overall direction of the information, application and operational security architecture as well as creating an understanding of risks by analysing and forecasting threats to information security. Moreover, the CISCO role includes managing security monitoring, analysis, detection and going through incident response processes fostering information security awareness and promoting a culture of information security and privacy.

Document All Changes. A change log document helps in keeping track of every change on productive systems for traceability and clearness. The service provider needs physical and IT security in place to defend the data centre from intruders. Therefore centre personnel need to be trained regularly. Part of these controls, are access control of the data centre and video surveillance of the server rooms. It must be clear, who has had access to which server rack at which time. Normally, video recording tapes are digitally stored for 24 hours and when there is no reported incident, they are erased.

Plan on Business Continuity and Disaster Recovery. The service provider's data centre needs physical and logical redundancy for all IT equipment (software and hardware) as well as a disaster recovery and business continuity plan for the data

centre. The service provider should formulare a disaster recovery plan together with partners, at least once a year. A thorough documentation of all relevant security processes builds the basis for an ISMS that supports the execution of the security policy.

Ensuring effective risk management Identify the information assets within the company that need to be protected and conduct an accurate risk assessment on them. With risk management, the risk can be handled or even accepted, when the costs for preventive measurements are too high. Before conducting a risk assessment it is well known to first classify the information assets based on their relevance and company impact. The assessor must know if their Confidentiality, Integrity or Availability can be compromised and to which extent.

Produce a control list consisting of action items The controls are made up of the risk assessment again and the overall approach for mitigating or leaving risks. Selected controls should be mapped to Annex A of the standard, which consists of 133 controls in 11 domains. A review of Annex A is used to figure out, if any control areas have been missed out in the compliance process.

4 A Mapping between ITIL and ISO 27001 Action Items

The two standards, ITIL and ISO 27001 are capable of mappings between them. We are using steps from the ITIL process as input for the ISO 27001 process. With these inputs, an output can be generated. Which means that the result of the ITIL process serves as input for the ISO 27001 activities. There are different action items regarding countermeasures within the ITIL and ISO 27001 standards. So how can the company compliance still be assured after the mapping between the ITIL and ISO 27001 processes took place? It is a matter of compliance regarding the privacy and legislation of the involved participants and entities. ITIL and ISO 27001 have in common that they are both based on the Plan-Do-Check-Act (PDCA) model. From ITIL point of view, nearly all security controls in ISO 27001 are part of ITIL service management. It is also in the ITIL standard, chapter on service design that there is a reference on ISO 27001. The advantage of this approach is that the company's information security department is in line with the risk management department regarding which ITIL processes have been implemented through ISO 27001. The information security department can also easily identify which ISO 27001 objectives are already met through the use of ITIL and which still must be handled. Using this hybrid approach, considering ITIL and ISO 27001 together, companies can save a lot on time and money using mutual synergies and knowledge in this area. The need to use both standards at the same time is to establish a well-known information security process that covers all relevant aspects. If doing so, a company can rely on acceptable security levels, effectively manage risks and reduce overall risk levels. We propose a mapping between ITIL and ISO 27001 action items based on the points in Sect. 3.1 and Sect. 3.2. Point 1 of ITIL, Define initial tasks for the roles is mapped to point 1 of ISO 27001, named Agree on an IT security policy. So starting with the initial tasks and roles, a security policy can be written and agreed on. The creation of an incident response system, point 2 of ITIL is the basis for establishing the position of a CISO in a company. So it is technical means, which is the

base for organizational duties of a CISO position. Moreover, step 1 of ISO 27001, the establishment of a security policy is needed to have a CISO on duty. So there are not only relations between the standards but also within one standard. Point 3 of ITIL, the establishment of a monitoring system, serves as input for action item 5 of ISO 27001, ensuring effective risk management.

Only with a thorough configured monitoring and alerting system it is possible to conduct effective risk management. The goal is to mitigate risks and be aware of new risks. A classification of the monitored assets is important to easily identify the severity of an incident. Reaction times, escalation times, contact with partners and producers and time to recover are heavily dependant on the severity of an incident. The higher the severity is, the faster is the required reaction times, escalation times, communication with partners and producers and recovery time. Only if an organisation is certified in its core business areas, can it be able to embark upon business continuity and disaster recovery planning. Certified personnel as well as experience and knowledge in the matter of subjects are a plus. Input for this is the documentation of all necessary information and changes. The documentation must be extended and kept up to date at every chance. It is important to define boundaries to stakeholders and produce a control list consisting of action items as well. The boundaries to stakeholders serve as input for Annex A of the ISO 27001 standard.

5 High Availability Video Conferencing Service Provider

Our example is based on a high availability video conferencing service provider, who is supporting customers with their company video infrastructure. Thus it needs high availability and fast response times to the end client.

5.1 Instantiate the Cloud Pattern

To illustrate our approach, we use the example of a service provider. This company wants to offer video conferencing services to its customers. The video conferencing services are planned to be implemented via a cloud system.

Figure 2 shows an instance of the extended cloud pattern for the description in the following: The cloud provider is a company called Alpha. The main goal of the cloud provider is to maximize profit by maximizing the workload of the cloud. Therefore its sub-goals are to increase the number of cloud customers and their usage of the cloud, i.e. to increase the amount of data as well as the number and frequency of calculation activities. Fulfilling security requirements is only an indirect goal to acquire cloud customers and convince them to increase the subset of processes they outsource. The pool of the cloud is instantiated with Alpha's data centers that consist of servers, network, and virtualization software. The data centers are located in Germany and the US. Alpha uses the internal help desk of its data centre as cloud support and the internal data centre IT services unit as cloud administrator. Alpha implements the Alpha cloud table as cloud database and the Alpha hypervisor as cloud hypervisor. The cloud customer is instantiated with a technology company that plans to outsource the effected IT processes to the cloud to reduce costs and scale up their system for a broader number of end customers.

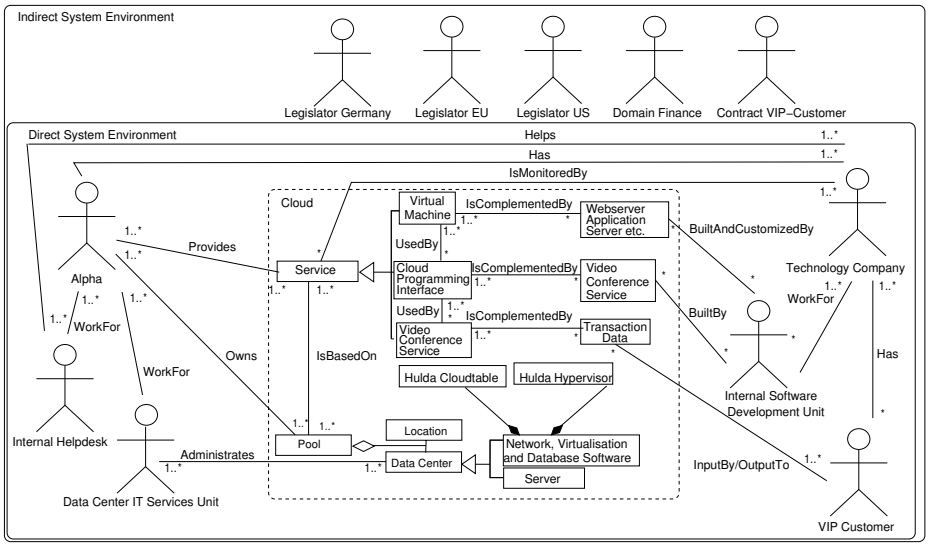


Fig. 2. Instantiated Cloud System Analysis Pattern with the Video Conferencing Scenario

Customer data such as IP addresses, names, and transaction log history are stored in the cloud. Transactions such as booked meetings are processed in the cloud. We instantiate the cloud developer with an internal software development unit of the technology company. This unit develops solutions for video conferencing in the cloud. The internal development unit installs and configures web- and application servers in virtual machines. The resulting cloud programming interface is the foundation for the video conferencing service. The technology company uses the cloud to conduct his/her financial business.

5.2 Describe Relevant Business Processes

The business process in our scenario is made up of a cloud customer, who demands the service of a video conferencing suite, especially made for VIP customers. To maintain such a high available solution, the cloud customer requests the following services from the cloud provider: A state of the art video conferencing system, using HD video quality and a fast and reliable connection between the video participants and conducting a penetration test against the video conferencing site once a year with a detailed report. The cloud customer network is maintained by the cloud provider, who is responsible for the security and non-disclosure of data. For HD video quality, the customer will need at least a 2 Mbit connection and preferably more bandwidth. It is also a good idea to have a single point of contact, globally reachable video support number that is used as the primary place to go for first level incidents. It is of course a good idea to have the WAN connections and the video conferencing service from the same cloud provider. There are not many providers who can offer both services on a worldwide basis. Once a week there needs to be ongoing discussion between the cloud provider and the customer concerning new video system installations and the incidents recorded at the first level

support. The cloud provider is also responsible for keeping up maintenance and service contracts of the video endpoints. If the first level support is not successful or it takes too long a time to recover, the case is escalated to second and third level. The second level can be within the cloud provider or local IT at the customer site. The third level is the producer of the video equipment, which can be a partner of the cloud provider and/or customer. The customer needs proofed processes in place for scenarios, such as a video conference room installation or the hardware replacement of a video endpoint.

5.3 Describe Services in Detail

In our example, the video conferencing service is provided to the customers by the cloud provider. It consists of the dial-in bridge and one-to-one movie connections as well as security mechanisms. If a change occurs, such as that caused by the installation of a security update on the management server, customers are redirected to a backup system, which indicates that the main site is under maintenance and will be back in production soon. If an incident outside a maintenance window occurs, the responsible IT manager of the cloud customer will contact the service delivery manager of the cloud provider. The given incident has to be handled and restored to normal working conditions in accordance to the time frames specified in the agreed Key Performance Indicators (KPIs). If this is not possible, the cloud provider has to pay a penalty for the non delivery of the managed service. For the compliance with the KPIs, service reports with the availability expressed in percentage of the provided service are delivered to the customer. To further analyze the quality of the service, end-to-end monitoring from the customer network to the login server can be applied. The end-to-end monitoring reports can be accessed from the cloud customer via a separate web page. The task of the end-to-end monitoring is to supervise the availability and response time of the login service and alert the provider as well as the customer in case of exceeding defined thresholds. The system further distinguishes between warnings and errors and is capable of alerting status changes. The end-to-end monitoring procedure stimulates the user experience and does so by executing a macro of user transactions constantly 5 times per hour. The security requirements for the retrieval of these monitoring reports involve the use of certificates and encryption as well as a complex algorithm used for client authentication. This algorithm shall consist of a combination of 3 different authentication mechanisms. First, the unique username, second user input by finishing a capture picture and third a unique password corresponding to the user. The certificate in use must be trusted by an authorized third party company, who issues the certificate and have an encryption standard of 256 bits. For VIP customers immersive, telepresence video rooms are used. This needs the involvement of different parties at the customer side, for example facility (cooling, heating, electrician,...) and the local video team. Once, all physical tasks have been carried out, the video system can be certified, prior to using it. The certification takes place between the customer site, who initiates a connection to the cloud provider and vice versa. Parts of this test includes presentation sharing, movement of people, taking a screenshot and having a stable connection for at least 15 minutes. The quality of the connection can be seen during the connection in the quality details. The technicians should have a look at parameters, such as delay, jitter and packet loss. A common

issue can also be one way audio or video, which is mostly because of missing firewall configuration on one participant side.

5.4 Instantiate ISMS Policies

First an assessment is conducted by the cloud customer in order to select a cloud provider that best fits his requirements. Normally such a selection is done through a bidding procedure. The cloud customer is describing the services he wants to consume as well as availability values and response times for incidents. The cloud customer should also check if the cloud provider has a qualified 24x7 support team in place, which the customer can call, when there is a major incident or outage. Another important point is to clarify if the provided infrastructure is solely in use for the customer and not part of a multi tenant solution. Further, disaster recovery on the hardware side (e.g.: VMware HA, VMware Vmotion) and the existence of a fallback data centre should be clarified. It is also interesting in which country the data centre resides and where the customer data is actually stored. This is important as which national law applies and where the court jurisdiction resides. Which security mechanisms does the cloud provider have in use and which quality of cooling, electricity venues are in the field. It should also be clear, if the cloud provider will monitor the customer's servers and be responsible for data backups. Generally speaking, the service quality should be constantly monitored by the cloud provider. This can for instance be done with an end-to-end monitoring solution, where end user behavior is simulated. Such a solution measures the availability from the customer point of view as well as response times and network monitoring. End-to-end monitoring solutions are collating these KPIs through measuring scripts running on separate machines crawling through the productive environment. The monitoring results are then reported to the cloud customer once a month for every single service in use. The measurement criterion for the service delivery and maintenance duties are the KPI and are based on the overall availability, availability of customer processes and customer satisfaction. The monitoring results are to be delivered to the customer for quality assurance at the beginning of the month. For service delivery, references of customers (success stories), partner management (which partners support the cloud provider) and a sample contract can be requested. The contract duration, opt-out criterion as well as price models for the services are necessary.

5.5 Conduct an Internal and External Security Audit

The cloud provider of course will get the chance to improve their offer in the second phase. A feedback process from the cloud customer to the cloud provider after the first phase is recommended. The services provided to the customer are defined in a contract between the two parties involved. In this contract the service levels are expressed in SLA, as well as the response times during business hours and after business hours are defined. A high customer satisfaction rate is assured through security, availability and confidence in the cloud provider. The cloud provider has established processes and policies that maintain the security of the customer data. A confident handling of customer information and infrastructure is based on the consequent implementation of rules and processes, which are embedded in policies and specified in standards and operational

guidelines. Implementing these guidelines is the duty of the cloud providers' employees. Based on these policies, responsibilities, roles, behavior, process definitions and supporting technologies are derived.

5.6 Implement Change Management

For the change log within the service delivery, all changes in the productive environment must be logged. The structure of the change log is as following: The first column is filled with an incremental ID, followed by the name of the change owner. The initiator of the change, which can vary, comes next. After that, the change cause, described in words or referring to a support ticket is listed. Next is the hostname, where the change is carried out, following the change description in descriptive language as well as the configuration changes itself in computer language. The category of the change is classified in regular, emergency and standard. It has to be defined, which duties belong to a standard change. If a change is non-standard, it is qualified by the engineer as a non-standard change. All productive changes have to be scheduled and appropriate configuration and fallback procedures in place prior to executing the change. The last possibility is a non-operational relevant change. This means changes in non productive environments and no scheduling of the change is needed. The risk of the change can vary from low to medium to high. At the end, a control check is performed, where the name of the change approver as well as the date the change was approved and the date the change was carried out, are documented. Last, but not least, a flag is inserted stating whether the change was successful or not. Changes are executed, following a four eye principle.With this feature, errors can be minimized and the overall availability and customer satisfaction is high. Urgent changes or emergency changes can be delivered faster, but require the consent of the Change Advisory Board (CAB).

6 Related Work

Calder [12] and Kersten et al. [13] provide advice for an ISO 27001 realisation. In addition, Klipper [14] focuses on risk management according to ISO 27005. This works do not consider relations to other standards like ITIL.

Cheremushkin et al. [15] [16] present a UML-based metamodel for several terms of the ISO 27000, e.g., assets. These meta-models can be instantiated and, thus, support the refinement process. However, the authors do not present a holistic approach to information security. The work mostly constructs models around specific terms in isolation. We support the establishment of an ISO 27001 ISMS using the processes in the ITIL standard.

Mondetino et al. investigate possible automation of controls that are listed in the ISO 27001 and ISO 27002 [17]. Their work can complement our own.

Fenz et al. [18] introduce an ontology-based framework for preparing ISO/IEC 27001 audits. They provide a rule-based engine which uses a security-ontology to determine if security requirements of a company are fulfilled. This work also does not consider the ITIL standard.

7 Conclusion and Future Work

We have established a method that helps with mapping ITIL action items to ISO 27001 action items. The method works in a compliant way, also after this mapping has been done. Our contribution is intended to help organisations, which are already certified or are thinking becoming certified.

Our approach offers the following main benefits:

- A structured method to map ITIL action items with ISO 27001 action items
- Systematic identification of relevant action items and determining compliance mechanisms for them
- Improving the outcome of business processes by adding benefits from mapping between standards
- Re-using the structured techniques of ITIL and ISO 27001 for supporting business processes in order to be compliant with current legislation and demands of law

As standardization is not an easy step for companies, we came up with suggested steps to help companies become certified. We presented a combination of technical and organizational means to master this process. Our cloud analysis pattern is the basis for the high availability video conferencing solution operating in the cloud. We apply our pattern on both, ITIL and ISO 27001 standards. Our hybrid approach, using the synergies from both certifications is shown in the mapping between them. In our example we present the whole life-cycle, the instantiation of the cloud pattern, the description of the relevant business processes, the description of the managed service in detail, the instantiation of ISMS policies, conducting internal and external securtity audits and finally, the implementation of change management within the organisation.

The work presented here is based on [5] and is further extended with the use of ITIL and our cloud- specific analysis pattern. Future work will involve the mastering of security and privacy questions within this context. The legislative framework in connection with both the cloud provider and the customer side must be examined in more detail. We think that this paper builds a stable basis for all of these challenges.

References

1. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I., Zaharia, M.: Above the clouds: A berkeley view of cloud computing. Technical report, EECS Department, University of California, Berkeley (2009)
2. Mell, P., Grance, T.: The NIST definition of cloud computing. Working Paper of the National Institute of Standards and Technology (NIST) (2009)
3. Vaquero, L.M., Rodero-Merino, L., Caceres, J., Lindner, M.: A break in the clouds: Towards a cloud definition. Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review 39(1), 50–55 (2008)
4. Buyya, R., Ranjan, R., Calheiros, R.N.: Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities. In: Proceedings of the International Conference von High Performance Computing and Simulation (HPCS). IEEE Computer Society Press (2009)

5. Beckers, K., Kuester, J., Faßbender, S., Schmidt, H.: Pattern-based support for context establishment and asset identification of the iso 27000 in the field of cloud computing. In: Proceedings of the International Conference on Availability, Reliability and Security (ARES). IEEE Computer Society Press (2011)
6. Jackson, M.: Problem Frames. Analyzing and structuring software development problems. Addison-Wesley (2001)
7. Fowler, M.: Analysis Patterns: Reusable Object Models. Addison-Wesley (1996)
8. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley (1994)
9. Schumacher, M., Fernandez-Buglioni, E., Hybertson, D., Buschmann, F., Sommerlad, P.: Security Patterns: Integrating Security and Systems Engineering. Wiley (2006)
10. International Organization for Standardization (ISO), International Electrotechnical Commission (IEC): Information technology - Security techniques - Information security management systems - Requirements. ISO/IEC 27001 (2005)
11. Government, H.: It infrastructure library (itil) (2012), http://www.itil-officialsite.com/home/home.aspx
12. Calder, A.: Implementing Information Security based on ISO 27001/ISO 27002: A Management Guide. Haren Van Publishing (2009)
13. Kersten, H., Reuter, J., Schroeder, K.: ITSicherheitsmanagement nach ISO 27001 und Grundschutz. Vieweg+Teubner (2011)
14. Klipper, S.: Information Security Risk Management mit ISO/IEC 27005: Risikomanagement mit ISO/IEC 27001, 27005 und 31010. Vieweg+Teubner (2010)
15. Cheremushkin, D., Lyubimov, A.: An application of integral engineering technique to information security standards analysis and refinement. In: SIN 2010 (2010)
16. Lyubimov, A., Cheremushkin, D., Andreeva, N., Shustikov, S.: Information security integral engineering technique and its application in isms design. In: Proceedings of the International Conference on Availability, Reliability and Security, ARES (2011)
17. Montesino, R., Fenz, S.: Information security automation: how far can we go? In: Proceedings of the International Conference on Availability. IEEE Computer Society Press (2011)
18. Fenz, S., Goluch, G., Ekelhart, A., Riedl, B., Weippl, E.: Information security fortification by ontological mapping of the iso/iec 27001 standard. In: Proceedings of the International Symposium on Dependable Computing. IEEE Computer Society Press (2007)

Towards Improved Understanding and Holistic Management of the Cyber Security Challenges in Power Transmission Systems

Inger Anne Tøndel[1], Bodil Aamnes Mostue[2], Martin Gilje Jaatun[1], and Gerd Kjølle[3]

[1] SINTEF ICT, Trondheim, Norway
inger.a.tondel@sintef.no
[2] SINTEF Technology and Society, Trondheim, Norway
[3] SINTEF Energy Research, Trondheim, Norway

Abstract. Information and Communication Technology (ICT) is increasingly utilised in the electrical power transmission system. For the power system, ICT brings a lot of benefits, but it also introduces new types of vulnerabilities and threats. Currently the interdependencies between the power and ICT system are not fully understood, including how threats (both malicious and accidental) towards the ICT system may impact on power delivery. This paper addresses the need for improved understanding between ICT security and power experts. It explains important terms used differently in the two disciplines, identifies main impacts on power systems that may result from ICT incidents, and proposes a set of indicators that can be used as a basis for selecting measures.

Keywords: information security, cyber security, power transmission, indicators.

1 Introduction

The next-generation electric power system (the smart grid) is central in dealing with emerging energy challenges. In Europe, the goal of 20 % improvement in energy efficiency, 20 % share of renewable energy, and 20 % reduction in greenhouse gas emissions by 2020 is a driver towards a smarter grid where renewable energy can be more effectively introduced, failures can be more easily detected and managed and where customers and their smart appliances consume energy in a more flexible way.

The electric power system can be divided into three main areas: *Generation* of energy, *transmission* of energy at high voltage over long distances, and *distribution* of energy at lower voltage towards customers. The transmission system is already quite smart compared to the distribution system, and most of the smart grid initiatives thus introduce more intelligence into the distribution grid. Advanced Metering Infrastructure (AMI) is one such modernisation.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 240–255, 2013.

For power systems, the N-1 principle has guided the work on securing the infrastructure; the power system should withstand loss of any single principal component without causing interruptions of electricity supply. Today however, N-1 security is challenged, mainly due to reluctance towards new power lines, massive integration of renewable energy sources, and the liberalisation of the electricity market. To deal with this, more intelligence has been introduced into the power system. something that has increased the complexity of the power system [1]. At the same time as N-1 security is challenged, it is also clear that N-1 security is not enough to prevent major events in the power system, as clearly demonstrated by recent blackouts such as in the US/Canada, Sweden/Denmark and Italy in 2003 [2], and the Europe blackout in 2006 [3].

To be able to secure the power system in a cost-effective way, it is important to understand where the system is most vulnerable and what measures will have most effect. Kröger and Zio [4] have provided an overview of approaches for vulnerability assessments of critical infrastructure. All approaches listed require quite detailed knowledge of the system. This is also the case for the quantitative analytical approach commonly taken for risk analysis of electricity supply [5], and for probabilistic modelling that have been pointed out by some as a way to increase cost-effectiveness in the security work [1, 6]. The deep knowledge of the electricity system, detailed computer models and specialized computer tools required makes such approaches difficult to use when also introducing "new" aspects, such as ICT. The models of the power system are already complex, and ICT components are present all the way from the bays to the control centres, constituting a big distributed ICT system. Creating a unified model that is reasonably correct is not easy. There is also a need to estimate failure probabilities of ICT components. Achieving confidence in such estimates is challenging as estimates will have to be made with limited experience data available. Some of the ICT components are relatively new, the introduction of ICT happens fast, and the threat landscape is constantly changing.

Because of the above challenges, there is a need for approaches that can add value also without excessive modelling effort and that can improve understanding of the interdependencies between power and ICT. To that end, this paper:

- contributes to increased mutual understanding between ICT and power experts
- identifies indicators that can be used to monitor trends when it comes to the risk of ICT security incidents in power systems

Currently the interdependencies between the power and ICT system are not fully understood, including how threats (both malicious and accidental) towards the ICT system may impact on power delivery [7, 8]. As a step towards increased understanding of these interdependencies, this paper address cyber security challenges of the transmission system, considering the ICT technology that can be found in such systems today. Though the focus is on the transmission system, many of the issues identified are likely to be also relevant for smarter distribution systems. The use of indicators can increase the Transmission Service Operator's (TSO's) understanding of the effects ICT have on power system security. This

will improve the basis for current decision making at TSOs, especially when it comes to measures needed, and can improve foundations for detailed vulnerability assessments later on.

The paper is organised as follows. Section 2 provides a brief overview of ICT components that are used in power transmission systems today. Section 3 provides an overview of central terms, with a main emphasis on terms that are used differently among ICT and power system experts. Section 4 identifies threats towards ICT systems that may impact power delivery. Section 5 gives an introduction to the role of indicators. Section 6 suggests and evaluates candidate indicators that can be used as a starting point by TSOs. Section 7 discusses the contribution of the paper, and Section 8 concludes the paper.

2 Overview of Main ICT Components in Power Transmission Systems

The ICT systems used for transmission [9–13] include monitoring, control and regulation systems, protection systems, defence systems, automation systems and communication systems. The transmission system itself is highly distributed and complex. The ICT system is also distributed and dependent on efficient and reliable communication technology. Communication equipment and critical ICT components are usually duplicated.

Wei et al. [13] divide the major functions of power grids into three levels: the corporate level, the control centre level and the substation level. The corporate level is concerned with business and operation management (more long term), while the control centre level and the substation level are involved in the real-time management of the power system. Important functions of the control centre level is forecasting of load and power generation sources, monitoring of system state, operation, system analysis, making recommendations, processing of alarms, training, logging and data exchange. At the substation level, the main functions are to perform normal operation (collecting data and alarms, sending them to the control centre, and executing commands from the control centre), exchange of protection data within the substation, emergency operation, engineering, logging and maintenance.

Typically, the power system is organised in a hierarchical fashion, where substations are under the control of a control centre, that again may be under the control of higher-level control centres [13]. This way you end up with Regional Control Centres (RCCs), National Control Centres (NCCs), and even also transnational control centres. Status information is measured at bay level, collected at substations, and then forwarded to the Supervisory Control and Data Acquisition (SCADA) system. State estimators make the system able to cope with missing or erroneous system variables [11]. The SCADA system and EMS (Energy Management System) provide operators with updated status information, and also the possibility to issue commands. Essential in this respect is the communication infrastructure connecting the control centres with the substations and their bays. The substations also have control rooms with monitoring and control capabilities.

In addition to the centralised control scheme, the transmission system also consists of a large number of distributed autonomous intelligent devices that take part in ensuring safe and secure operation of the transmission grid. The most important example is represented by the protection devices [14] that have an important role in ensuring any failures have as little impact as possible, e.g. by disconnecting faulty parts of the grid. Protection devices come in different types. Many of them require communication in order to work effectively, and have strict requirements when it comes to response times. Protection devices are usually duplicated at the transmission level.

Lately, more intelligence has been introduced into the power system in form of Special Protection Schemes (SPSs) that are able to implement corrective actions automatically and cover a wider area, and Phase-Shifting Transformers (PSTs) and Static VAR Compensators[1] (SVCs) that increase controllability [1]. Wide-Area Monitoring Systems (WAMS) provide enhanced situational awareness and thus assist in decision making at control centres [11].

It is worth noting that a lot of the components and systems mentioned above, like protections, are not traditionally considered to be ICT. Still, the current development has turned also these devices into small computers that rely on communication. In this paper they are thus considered part of the ICT system.

3 Terms: Communication Challenges between ICT and Power Experts

Properly addressing the ICT risks of the combined ICT and transmission system requires cooperation between people from different disciplines: experts on electric power systems and experts on ICT, people with dependability background, people working on safety, and people with cyber security background. The terms used vary between the different disciplines, with one of the most confusing terms being "security".

3.1 Security, Information Security and Cyber Security

In the context of ICT, the work on protecting the systems is usually denoted information security. The key asset to protect is considered to be the information in the systems, and the goal of the information security work is to ensure [15]:

- Confidentiality: *"the property that information is not made available or disclosed to unauthorized individuals, entities, or processes"* [15]
- Integrity: *"the property of safeguarding the accuracy and completeness of assets"* [15]
- Availability: *"the property of being accessible and usable upon demand by an authorized entity"* [15]

[1] Volt-Ampere Reactive (VAR) compensators control reactive power injection or absorption in order to improve the performance of the transmission system [11].

The terms computer security and (ICT) network security[2] are also sometimes used, and when ICT is used in critical infrastructure, the term cyber security is common. The difference between information security, computer security, network security and cyber security is not clearly defined. It can be argued that information security is a broader term as it includes also information not processed, stored or communicated by means of ICT. Computer security and (ICT) network security is more centred on protecting (specific parts of) the ICT technology. Cyber security is focused on the threats coming from closer integration with the Internet, and seems to be primarily used when talking about the security of industrial control systems. The terms are however often used interchangeably, and their definitions are often quite similar. As an example, the NISTIR 7628 guidelines on smart grid cyber security [16] explain cyber security in terms of ensuring confidentiality, integrity and availability of electronic information communication systems. For power systems, however, the term security has a different meaning. According to Kundur et al. [17], *"Security of a power system refers to the degree of risk in its ability to survive imminent disturbances (contingencies) without interruption of customer service."*

3.2 Dependability of ICT Systems vs. Information Security

Dependability of ICT systems can be defined as *"the ability to deliver service that can justifiably be trusted"*, or alternatively *"the ability to avoid service failures that are more frequent and more severe than is acceptable"* [18]. According to Avizienis et al. [18], dependability is comprised of the five attributes: availability, reliability, safety, integrity and maintainability. Thus, dependability and information security have two attributes in common: integrity and availability. These attributes are essential in automation applications. The confidentiality attribute of information security is however not considered for dependability. Traditionally, dependability has been concerned with non-malicious faults while information security has paid attention to the threats posed by malicious actors. There has however for a long time been a growing understanding among the dependability experts that limiting studies to non-malicious faults implies only addressing part of the problem. Among the information security experts, it is evident that also non-malicious faults can cause severe problems for the confidentiality, integrity and availability of information. This is reflected in standards such as ISO/IEC 27005 [19] on information security risk management that specifically states that analyses should include both natural threats and threats with human origin, and also both accidental and deliberate threats. In critical infrastructure protection, the term "all-hazard approach" is often used to emphasise the need to include both natural and man-made events, and also both intentional and unintentional actions [20].

The terms "hazard" and "threat" have similar meanings. The term *threat* is defined in ISO/IEC 27002 as *"a potential cause of an unwanted incident,*

[2] The term 'network' in network security' refers to the ICT network, not the power network.

which may result in harm to a system or organisation" [21]. The term *hazard* is defined in IEC 61508 as *"Potential source of harm"* [22]. The term hazard is more commonly used for safety and dependability analysis, whereas the term threat is more commonly used for information security[3].

4 ICT Threats Relevant for the Power System

One step on the way towards a better understanding of the vulnerability that comes from the deep integration of power and ICT systems is to identify the main incidents that may happen in the ICT system, and that may have consequences for power delivery. According to Doorman et al. [23], the major unwanted situations in the energy sector are:

- high price: the price of electricity is higher than usual for a long period
- load curtailment: rationing
- blackouts: interruptions for longer periods of time

As ICT is deeply integrated with the power system, ICT failures may have consequences also along these lines. This is supported by results from the GRID project that identified direct effects ICT system failures may have at the transmission grid level [7]. Among the effects identified were power system instability, loss of generation capacity and loss of lines and/or corridors, malfunction of protection devices or other devices used for power control, and loss of or corrupted observability (e.g. SCADA or EMS).

As outlined in Section 2, the ICT systems used for transmission include monitoring, control and regulation systems, protection systems, defence systems, automation systems and communication systems. According to a survey performed by Gardner et al. [24], industry and research communities agree that the most critical areas are protection and control. This is because of the role that protection and control systems play in normal and abnormal operation, and also the potential consequences of single errors in these types of systems.

The NERC Cyber Attack Force [25] identified a set of cyber attack scenarios for power systems. The cyber attack scenarios were intended to be used as a basis for operational training, and cover the following unwanted situations: social engineering [26] where a false request or false information is sent to an operator; denial of service of EMS network; denial of service of EMS applications; spurious device operations, and; realistic data injection. The cyber attack scenarios listed first are considered more plausible than those listed last.

The different ICT subsystems have different characteristics and may be subject to different types of attacks or failures. Potential threats and vulnerabilities of relevant systems have been documented in several publications. As examples, Wei et al. [13] have identified potential network attacks on typical communication links used for smart grid automation systems, and also the potential adverse impacts of such attacks. In addition, they have identified the main targets of attacks

[3] Note that some definitions of the term "threat" only include intentional and malicious acts [20].

on the SCADA system, including the Front End Processor (FEP), the Human Machine Interface (HMI), the Engineering Workstation (EWS), the database systems, the application server, and the controllers. The National SCADA TestBed (NSTB) [27] has, based on a number of security assessments of SCADA systems, shared information about what types of cyber security vulnerabilities are commonly found in SCADA systems. Sridhar et al. [11] have identified potential cyber vulnerabilities related to state estimation, VAR compensation and Wide-Area Monitoring Systems.

In the following we describe unwanted situations in the ICT system that may cause the effects listed above. The unwanted situations are related to the control and protection area (most critical), and take into account the need for operators to see (observability) what is happening and act (controllability) on what they see. They also consider the potential goal of attackers to influence the state of the system (command injection).

4.1 Loss of or Corrupted Observability

Loss of or corrupted observability happens in cases where operators, either at a regional or national control centre (RCC/NCC), do not have a correct overview of the current state of the system. This may be due to information not being available (loss of observability) or because of erroneous information in the system (loss of integrity). The consequences of unavailable or erroneous information for the power system can vary from no consequences to high consequences. The consequence is dependent on the degree to which observability is lost or corrupted, i.e. the duration of the incident as well as how much of the system is affected. It is also dependent on the state of the system at the time, and what happens while observability is lost or corrupted. The ability to detect corrupted observability is also important. Consequences may be higher if operators are not aware that information is corrupted, and thus may act on the erroneous state information. Loss of or corrupted observability may happen due to:

- Failure of communication equipment or noise on the communication channel, causing data transmission errors, unavailability of the communication line or excessive delays in the communication
- Failure of central systems at the RCC/NCC, causing the control centre to be unable to communicate or causing unreliable reception of status updates (random modifications)
- Failure of components at the substations, causing the substations to be unable to communicate or causing unreliable status measurements or unreliable processing of status updates at substations
- Attack on the communication ability between RCC and substations
- Injection of false status information

If loss of or corrupted observability is caused by deliberate attackers, this may intentionally hide other malicious activities in the system. Thus, operators are less able to detect the malicious activity and take action. In such cases, the consequences of lacking observability are likely to be higher than when caused

by natural failures. False status messages that have been deliberately injected into the system are also more likely to be specifically tailored to trigger specific actions. The total observation of the state of the system may also be changed in a way that seems convincing. Such deliberate corruption of observability by an intelligent adversary is thus more likely to cause major consequences than random failures. Manipulating status information undetected is however very difficult for an attacker, as state estimators verify that the current status information is coherent. To be able to change the state in a coherent way, attackers will require detailed knowledge of the current situations. In addition, attackers must be able to modify all signals and measurements in a coherent way.

4.2 Uncontrollability of the System

Uncontrollability of the system happens in cases where operators of the SCADA system are not able to send commands in order to control the power system. Uncontrollability also happens if the components that act on the commands stop responding to commands or responds in an unintended way due to malfunction or an attack. Also in this case, the consequences are dependent on the duration and extent of the uncontrollability, and the state of the system at the time. Loss of controllability may happen due to:

- Failure of communication equipment or noise on the communication channel, causing data transmission errors, unavailability of the communication line or too high delays in the communication
- Failure of central systems or components at the substations, RCC or NCC, causing the systems to be unable to communicate or causing unreliable sending or reception of messages
- Failure of components that act on commands
- Attack on the communication ability of central systems

If uncontrollability is caused by attackers, as opposed to random failures, the consequences are likely to be more severe as attackers are likely to cause uncontrollability of central components at a time when such control is needed.

4.3 Command Injection

Command injection happens if components receive commands that have not been sent by authorised operators. Commands may be injected by attackers that have gained access to system components or the communication channel, but may also happen due to malfunction of equipment. If attackers are able to inject commands that are acted upon (e.g. open/close breakers), they can cause a lot of damage. Though operators are trusted, it is also important to be aware of the potential of operators to cause harm intentionally or due to mistakes. Attackers may also cause injection of unnecessary and potentially harmful commands indirectly by tricking operators to take unnecessary actions. This may happen due to erroneous information in the system (see Section 4.1) or through social engineering attacks [25].

4.4 Protection System Malfunction

Protection systems are essential for power system security, and protection system malfunction may have severe consequences for the power system [28]. Protection system malfunction may be caused by, e.g.:

- Protection systems with incorrect settings, either by mistake or by intent
- No communication or excessive delays in communication between protection systems
- Spurious tripping (caused by technical fault or human error)
- Desynchronized protection systems
- Errors in protection communication, either due to failures or due to modification by a malicious attacker

It is important to be aware that the protection systems operate under strict time conditions, where fault detection, protection decision and isolating device operation must happen in the space of milliseconds [14].

5 An Introduction to Indicators

The potential unwanted situations described above motivate the need to properly include ICT in work on power system security. As pointed out in the introduction, this is however not trivial and there is a need for improved understanding on the role ICT actually plays and how to properly grasp this in power security analyses. Using indicators is one way to increase understanding.

An indicator is a measurable and operational variable that can be used to describe the condition of a broader phenomenon or aspect of reality [29]. The word indicator comes from the verb *indicate*, which means to designate or to show. An indicator gives a simplified signal of a condition or change in condition, and indicators are typically used when the phenomenon itself is too complicated or too expensive to measure directly. The typical properties of indicators are [30]:

- They provide numerical values (a number or a ratio).
- The indicators are easily updated at regular intervals.
- They only cover some selected determinants of overall safety, security or risk, in order to have a manageable set.

Both individual indicators and their combinations can be useful since one can create a simplified description of the vulnerability level in the system and assess the expected performance and its development. Indicators can be used to see the long-term evolution (trends over months or even years), but also to observe sudden changes. Indicators can be characterised as leading (proactive, i.e., things that happen before an incident) or lagging (reactive, i.e., things that happen as a cause of or after an incident).

Table 1 provide an overview of existing approaches for developing indicators. *Risk based methods* are based on the hypothesis that risk control can be achieved

Table 1. Overview of approaches to identify indicators

Approach	*Description*
Risk based methods [29, 31]	Utilize a risk model as a basis. This is usually an existing risk model, but the development of a risk model may also be part of the method.
Safety performance based methods [32–35]	Start from a set of influencing factors assumed to be important to safety.
Incident based methods [36]	Identifies indicators by an in-depth study of one or more incidents or accidents.
Resilience based methods [37]	Identify indicators of human and organisational performance in a resilience perspective

through the control of risk influencing factors (RIFs), which are those factors having effect on risk. The conditions for this hypothesis to be true are that 1) All relevant RIFs are identified, 2) The RIFs are measurable and 3) The relationship between RIFs and risk is known. *Safety performance* indicators are based on important factors, but they are not related to a risk analyses. This also applies to *incident based methods* and *resilience based methods*. The resilience based method differs from the others by focusing on positive signals ("what went right" and why), rather than failures ("what went wrong").

The methods differ in scope and depth of analysis. The risk based approaches will cover the whole installation and all risks, since these methods narrow the focus to the most important risk factors. The resilience based approach can in principle also cover a complete installation with all its risks. The performance based methods will usually narrow the scope to certain systems or activities. The incident based methods will usually only cover specific systems and not a complete installation, but may go deeper into an area or system, which the other methods perhaps will not cover. Also, with the risk based methods it is easy to determine the risk significance, including relative risk importance between the various risk influencing factors. A main weakness with the other methods is that the risk significance and the relative importance of influencing factors or causes are unknown.

The literature describes a number of desirable properties for an indicator; but in practice, it appears very challenging to find indicators that meet all the requirements which are desirable. Below we list a selection of criteria for good indicators [38–42]:

- *Relevance (meaning):* The indicator value is assumed to be (strongly) correlated with either event frequency or consequence; or possibly with influencing factors or important parameters of a risk or vulnerability model; e.g. with Risk Influencing Factors (RIFs) of an influence diagram.
- *Availability:* Data to calculate the indicator can be acquired at a reasonable cost.
- *Reliability:* Data measured is regarded as being objective and without significant sources of error.

- *Completeness:* The total set of indicators should be complete, i.e. should cover all major types of hazards and system vulnerabilities.
- *Ownership:* A sense of "ownership" should be instilled in the users of the indicators (this is hopefully a consequence of the other criteria).

6 Cyber Security Indicators for TSOs: An Initial Set

To assess the vulnerability of the complete power and ICT system, it would be beneficial to create a risk model of the power system incorporating the ICT aspects, where all risk influencing factors are identified and measureable, and the relationships between the risk influencing factors are known. Risk indicators can then be used to describe the condition of risk influencing factors. However, it is very challenging to achieve this, because of the complexity of the system and the dynamics of an incident; there are many potential courses of events following a single incident, and human and organisational factors play a vital part in deciding how the skein will unravel. Furthermore, new types of threats and attacks are constantly emerging, and there is a lack of historical experience data. It is a challenge to estimate the contribution of human and organisational factors to the risk in a system. Detailed risk calculations will give results with considerable uncertainty due to the uncertainty related to such risk modelling, estimation of risk influencing factors, the relationships between them and lack of data.

Due to the current limited understanding of how ICT influences power system security, we have chosen to use a more simple approach to identifying indicators in this work. The literature describes several metrics that can be useful to measure trends of risk influencing factors [37, 43–45]. Based on this literature we have selected a set of indicators which can be relevant to improve understanding of the risk related to ICT incidents for transmission systems. About half of the indicators are based on indicators suggested by Gelbstein [45] (I3-5, I8, I13-14, I16, I18-20 in Table 2) and the rest were identified during a workshop with four experts on ICT, power systems and the use of indicators (I2, I6-7, I9-11, I17) or in discussions with colleagues before the workshop (I1, I12, I15). In the workshop all identified indicators were assessed based on their relevance, availability and reliability.

The selection of indicators has the following limitations:

- The hazards/threats mainly cover malicious attacks (security), but dependability issues (hardware and software problems) are also considered.
- A variant of a safety performance based method is used in the development of indicators, i.e. identifying indicators from factors that influence the security based on literature and expert knowledge.
- The identified indicators must be considered as examples, not an exhaustive list.
- The indicators are numerical values, monitored automatically or easily updated at regular intervals.

The factors and conditions that are taken into account and that may influence the security of the transmission systems are: threats to the ICT system, unwanted events, the performance of barriers of the ICT system, availability of the ICT system, and resilience of the TSO's organisation. An overview of the indicators and their evaluation score, as regards relevance (C1), availability (C2) and reliability (C3), is shown in Table 2. Scores were given from 1-5, where 5 is best.

7 Discussion

Today, the electrical power system is strongly dependent on ICT. Common cause failures can affect both infrastructures due to location-specific and functional interdependencies. There is a need to analyse how power systems and ICT interact and depend on each other, and these insights are lost if the technologies from these two disciplines are only studied separately. For efficient cooperation between power system and ICT experts, it is however crucial to take into account the difference in culture and use of terms in the two disciplines.

The combined power and ICT system is complex and highly distributed. Existing methods for vulnerability analysis often require quite explicit and detailed knowledge of the system. Currently, such deep knowledge is not available when it comes to the interrelations between power systems and ICT. There is thus a need to improve the understanding of the role of ICT, including the potential consequences of unwanted ICT incidents when it comes to power delivery. In this paper we contribute towards this by providing an overview of potential ICT incidents that may have such consequences. We also propose an initial set of indicators that can be used to monitor and improve understanding of the security of the ICT components and their influence on power system security.

It is not possible to measure the safety, security or vulnerability of such a complex system directly. Indicators can be used to measure factors that are important for the vulnerability of a system, and particularly the trends associated with such factors. In a complex system, there are several such factors and thus a high number of potential indicators. Still, it is important to have a manageable set of indicators that can be regularly monitored and followed up. The indicators suggested in this paper represent a first step towards identifying relevant indicators for TSOs that wish to monitor ICT risks and their impact on power security. Further work should include further identification, evaluation and testing of indicators, and should examine more carefully the interrelations and coverage of the suggested indicators.

The usefulness of indicators depend on how they are used in the organisation. An indicator or a combination of indicators with values outside the predetermined acceptable limits usually require actions. Input from indicators can also be used in deciding whether additional measures are necessary.

Table 2. Overview of indicators; C1 = Relevance, C2 = Availability, C3 = Reliability

Influencing factor	*Indicator*	*C1*	*C2*	*C3*
Threats	I1: Measurement describing the network traffic	3	5	5
	I2: Number of events (e.g. number of malicious attacks) of a certain type during e.g. the last month	5	2	2
	I3: Number of attempted intrusions detected [45]	4	2	2
Unwanted events	I4: Number of successful intrusion detected [45]	4	4	5
Barriers	I5: Number of orphaned accounts for access to sensitive information or critical systems [45]	4	4	3
	I6: Percentage share of identified software vulnerabilities not patched	4	3	3
	I7: Number of former employees still having access	4	5	4
Availability	I8: Total downtime (per critical ICT system) in the period reported (planned and not planned) [45]	3	5	4
	I9: Total downtime in the power system due to ICT interruptions in the period reported (planned and not planned)	4	5	4
	I10: Number of interruptions with downtime in the ICT system (operation regularity)	3	5	4
	I11: Number of interruptions with downtime in the power system due to ICT (operation regularity)	5	4	3
Resilience	I12: Proportion of personnel (%) working in ICT-systems with formal expertise in ICT	4	4	3
	I13: Number of unfilled positions in the ICT organisation [45]	3	4	4
	I14: Number of "near misses" in information security activities where no incident occurred, but could happen with small changes in the events [45]	5	1	2
	I15: Time from a vulnerability is reported to feedback is given	1	4	3
	I16: Mean time required to close a reported critical security incident [45]	4	4	3
	I17: Number of reported critical security incident not handled (backlog)	4	4	4
	I18: Number of related critical audit recommendations that have not been implemented [45]	4	4	5
	I19: Number of high impact items registered where mitigation activities have not been completed [45]	4	4	4
	I20: Number of information security processes or activities carried out by a single individual for whom there is no immediate backup or replacement [45]	4	2	2

8 Conclusion

Terminology is a challenge in a multidisciplinary field, and best results are achieved when terms are explicitly defined. Many information security challenges can affect power transmission systems, and using indicators is a promising way to work proactively with information security in a traditionally safety-oriented domain.

Acknowledgments. The work has been done as part of the AFTER project funded by EU, grant. nr. 267188. The authors would like to thank the other partners in this project for their cooperation, and especially the Coordinator Emanuele Ciapessoni from RSE.

References

1. Panciatici, P., Bareux, G., Wehenkel, L.: Operating in the fog: Security management under uncertainty. IEEE Power and Energy Magazine 10(5), 40–49 (2012)
2. Andersson, G., Donalek, P., Farmer, R., Hatziargyriou, N., Kamwa, I., Kundur, P., Martins, N., Paserba, J., Pourbeik, P., Sanchez-Gasca, J., Schulz, R., Stankovic, A., Taylor, C., Vittal, V.: Causes of the 2003 major grid blackouts in north america and europe, and recommended means to improve system dynamic performance. IEEE Transactions on Power Systems 20(4), 1922–1928 (2005)
3. Union for the Coordination of Transmission of Electricity (UCTE): Final report - system disturbance on November 4 2006 (2007)
4. Kröger, W., Zio, E.: Vulnerable Systems, 1st edn. Springer Publishing Company, Incorporated (2011)
5. Kjølle, G., Gjerde, O.: Risk analysis of electricity supply. In: Hokstad, P., Utne, I.B., Vatn, J. (eds.) Risk and Interdependencies in Critical Infrastructures. Springer Series in Reliability Engineering, pp. 95–108. Springer, London (2012)
6. Ciapessoni, E., Cirio, D., Grillo, S., Massucco, S., Pitto, A., Silvestro, F.: Operational risk assessment and control: A probabilistic approach. In: Innovative Smart Grid Technologies Conference Europe (ISGT Europe), pp. 1–8. IEEE PES (2010)
7. The GRID consortium: ICT Vulnerabilities of Power Systems: A Roadmap for Future Research (2007)
8. Hokstad, P., Utne, I.B., Vatn, J. (eds.): Risk and Interdependencies in Critical Infrastructures – A Guideline for Analysis. Springer Series in Reliability Engineering. Springer
9. Egozcue, E., Rodríguez, D.H., Ortiz, J.A., Villar, V.F., Tarrafeta, L.: Smart Grid Security, Anex I. General Concepts and Dependencies with ICT. Technical Report Deliverable - 2012-04-19, ENISA (2012)
10. Wang, W., Xu, Y., Khanna, M.: A survey on the communication architectures in smart grid. Computer Networks 55(15), 3604–3629 (2011)
11. Sridhar, S., Hahn, A., Govindarasu, M.: Cyber physical system security for the electric power grid. Proceedings of the IEEE 100, 210–224 (2012)
12. MIT: The Future of the Electric Grid. An Interdisciplinary MIT Study (December 2011)
13. Wei, M.D., Lu, Y., Jafari, Skare, P.M., Rohde, K.: Protecting smart grid automation systems against cyberattacks. IEEE Transactions on Smart Grid 2(4), 782–795 (2011)

14. Mesbah, M., Samitier, C., Einarsson, T., Acacia, M., Alvarez, J., Carmo, U., Castro, F., Cimadevilla, R., Darne, J., Dollerup, S., Freitas, J., Komatsu, C., Leroy, T., Ordunez, M.A., Runesson, A., Spiess, H., Stockton, M., Struecker, A., Valente, M., Vianello, G., Viziteu, I., Wright, J.: Line and system protection using digital circuit and packet communication. Technical Report JWG D2B5.30, CIGRE (2012)
15. ISO/IEC 27001:2005 (Information technology - security techniques - information security management systems - requirements)
16. The Smart Grid Interoperability Panel - Cyber Security Working Group: NISTIR 7628: Guidelines for smart grid cyber security: Vol. 1, smart grid cyber security strategy, architecture and high-level requirements (2010)
17. Kundur, P., Paserba, J., Ajjarapu, V., Andersson, G., Bose, A., Canizares, C., Hatziargyriou, N., Hill, D., Stankovic, A., Taylor, C., Cutsem, T.V., Vittal, V.: Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions. IEEE Transactions on Power Systems 19(3), 1387–1401 (2004)
18. Avizienis, A., Laprie, J.-C., Randell, B., Landwehr, C.: Basic concepts and taxonomy of dependable and secure computing. IEEE Transactions on Dependable and Secure Computing 1(1), 11–33 (2004)
19. ISO/IEC 27005:2008 (Information technology - Security techniques - Information security risk management)
20. Zio, E., Piccinelli, R., Sansavini, G.: An All-Hazard Approach for the Vulnerability Analysis of Critical Infrastructures. In: Proceedings of the European Safety and Reliability Conference 2011, Troyes, France, pp. 2451–2458 (September 2011)
21. ISO/IEC 27002:2005 (Information technology - security techniques - code of practice for information security management)
22. IEC 61508 (Functional safety of electrical/electronic/programmable electronic safety-related systems)
23. Doorman, G.L., Uhlen, K., Kjølle, G.H., Huse, E.S.: Vulnerability analysis of the nordic power system. IEEE Transactions on Power Systems 21(1), 402–410 (2006)
24. Gardner, R., Consortium, G.: A survey of ICT vulnerabilities of power systems and relevant defense methodologies. In: Power Engineering Society General Meeting, pp. 1–8. IEEE (2007)
25. NERC: Cyber Attack Task Force, Final Report (2012)
26. Orgill, G.L., Romney, G.W., Bailey, M.G., Orgill, P.M.: The urgency for effective user privacy-education to counter social engineering attacks on secure computer systems. In: Proceedings of the 5th Conference on Information Technology Education, CITC5 2004, pp. 177–181. ACM, New York (2004)
27. National SCADA Test Bed (NSTB): Common Cyber Security Vulnerabilities Observed in Control System aSsessments by the INL NSTB Program. Technical Report INL/EXT-08-13979, Idaho National Laboratory (2008)
28. Kjølle, G.H., Gjerde, O., Hjartsjø, B.T., Engen, H., Haarla, L., Koivisto, L., Lindblad, P.: Protection system faults – a comparative review of fault statistics. In: International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2006, pp. 1–7 (2006)
29. Øien, K.: Risk indicators as a tool for risk control. Reliability Engineering & System Safety 74(2), 129–145 (2001)
30. Øien, K., Utne, I.B., Herrera, I.A.: Building safety indicators: Part 1 theoretical foundation. Safety Science 49(2), 148–161 (2011)

31. Vinnem, J.E., Bye, R., Gran, B.A., Kongsvik, T., Nyheim, O.M., Okstad, E.H., Seljelid, J., Vatn, J.: Risk modelling of maintenance work on major process equipment on offshore petroleum installations. Journal of Loss Prevention in the Process Industries 25(2), 274–292 (2012)
32. UK Health and Safety Executive (HSE): Development process safety indicators. a step-by-step guide for chemical and major hazard industries (2003)
33. Centre for Chemical Process Safety (CCPS): Process safety leading and lagging metrics. you dont improve what you dont measure (2008)
34. Organisation for Economic Cooperation and Development (OECD): Guidance on developing safety indicators related to chemical accident prevention, preparedness and response. OECD Environment, Health and Safety Publications. Series on Chemical Accidents, 19 (2008)
35. Electric Power Research Institute (EPRI): Final report on leading indicators of human performance (2001)
36. Øien, K.: Development of early warning indicators based on accident investigation. In: PSAM9 International Probabilistic Safety Assessment and Management Conference (May 2008)
37. Øien, K., Massaiu, S., Tinmannsvik, R., Strseth, F.: Development of early warning indicators based on resilience engineering. In: PSAM10 International Probabilistic Safety Assessment and Management Conference (June 2010)
38. Rockhwell, T.H.: Safety performance measurement. Journal of Industrial Engineering 10, 12–16 (1959)
39. Kjellén, U.: The safety measurement problem revisited. Safety Science 47, 486–489 (2009)
40. Kjellén, U.: Prevention of Accidents through Experience Feedback. Taylor & Francis, London (2000)
41. Vinnem, J.E.: Risk indicators for major hazards on offshore installations. Safety Science 48(6), 770–787 (2010)
42. Herrera, I.A., Hollnagel, E., Håbrekke, S.: Proposing safety performance indicators for helicopter offshore on the norwegian continental shelf. In: 10th International Probabilistic Safety Assessment & Management Conference(PSAM 2010), Seattle, USA (2010)
43. SANS Institute: Twenty Critical Security Controls for Effective Cyber Defense: Consensus Audit Guidelines (CAG), version 3.1 (October 2011)
44. Herrmann, D.S.: Complete Guide to Security and Privacy Metrics. Measuring Regulatory Compliance, Operational Resilience and ROI. Auerbach Publications, Taylor & Francis Group, New York (2007)
45. Gelbstein, E.E.: Designing a Security Audit Plan for a Critical Information Infrastructure (CII). In: Laing, C., Badii, A., Vickers, P. (eds.) Securing Critical Infrastructures and Critical Control Systems: Approaches for threat Protection, pp. 262–285. IGI Global (2013)

Seeking Risks: Towards a Quantitative Risk Perception Measure

Åsmund Ahlmann Nyre[1,2] and Martin Gilje Jaatun[2]

[1] Norwegian University of Science and Technology
Department of Computer and Information Science
Trondheim, Norway
[2] SINTEF ICT
Trondheim, Norway
{asmund.a.nyre,martin.g.jaatun}@sintef.no

Abstract. Existing instruments for measuring risk perception have focused on an abstract version of the concept, without diving into the the details of what forms the perception of likelihood and impact. However, as information security risks become increasingly complex and difficult for users to understand, this approach may be less feasible. The average user may be able to imagine the worst case scenario should an asset be compromised by an attacker, but he has few means to determine the likelihood of this happening. In this paper we therefore propose a different approach to measuring risk perception. Based on well established concepts from formal risk analysis, we define an instrument to measure users' risk perception that combines the strengths of both traditional risk perception and formal risk analysis. By being more explicit and specific concerning possible attackers, existing security measures and vulnerabilities, users will be more able to give meaningful answers to scale items, thereby providing a better and more explanatory measure of risk perception. As part of the instrument development we also elaborate on construct definitions, construct types and the relationship between these and the corresponding risk perception instrument. Although it remains to be verified empirically, the validity of the measure is discussed by linking it to well established theory and practice.

1 Introduction

There is a fundamental relationship between risk exposure and the perceived need for protection. If there are no risks, then there is nothing that needs protection either. That is why users' intention to adopt a security measure is tied directly to the perceived risk [23]. However, when faced with

However, the problem is that perceived risk is highly subjective, and therefore varies greatly between people. There are many risk assessment methodologies, but as a rule they are quite complex, and not suitable for use by a layperson. In order to measure a layperson's risk perception, we need a simpler instrument that helps splitting the difficult question "what is the risk to your system" into

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 256–271, 2013.

manageable pieces. Even more importantly, though, the instrument should help to explain *why* the risk is perceived as it is.

A better understanding of regular users' risk perception would be an important basis for improving security technology, awareness-raising and ultimately also the uptake of security technology. This user-centric approach to risk allows security technology to focus on the risks important to the user, and thereby both gaining acceptance and motivation for its usage. The approach lends its idea from user-centric design of software systems, in that it is the users' perception that guides the design and presentation, rather than the designer's perception. This is not to say that security professionals' risk analysis should be discarded, but rather that the presentation to users should be based on their perceived risk. With a deeper understanding of the motivations behind the risk perception, it is also easier to spot misconceptions which can skew the perceived risk.

In order to gain a better understanding of risk perception, we need a model of how risk perception is formed and how it can be measured. This paper contributes a theoretical foundation and the initial step towards such a measurement instrument. To this end, we combine results from research on risk perception and well established concepts from formal risk analysis methodologies.

The remainder of this paper is organised as follows. In Section 2 we provide an overview of related work on measuring risk perception in IT. Next, in Section 3 we provide some background on previous research on judgments under uncertainty, risk perception, and risk analysis frameworks. In Section 4 we present our risk perception model and a preliminary instrument for measuring risk perception. Then we discuss the validity of our instrument in Section 5, before we give our concluding remarks in Section 6.

2 Related Work

Work on risk perception has been seen when investigating the factors that affect IT technology adoption in general, and information security technology in particular. For instance Featherman and Pavlou [8,25] showed that risk perception have a negative effect on consumers' intention to engage in online shopping. Risk perception was measured by the different facets of risks, including time risk, performance risk, financial risk and privacy risk. According to the definitions of these terms, they do not explicitly include the risks associated with active attackers. In a similar study, Kim, Ferrin and Rao [18] used the perceived privacy and security protection as well as familiarity, reputation, trust disposition and the presence of privacy seals to predict intention to adopt online shopping. The measure of privacy protection does include a scale item concerning "unautorized users (i.e. hackers)" [18], but for the most part the measure concerns the integrity and honesty of the vendor. Hörst, Kuttschreuter and Gutteling [12] and Belangr and Carter [1] both used a measure of risk perception in determining the intention to adopt e-government services. Both studies use a high level measure of perceived risk that neither caters for the different facets of risks nor active attackers.

Studies of adoption of information security technology such as anti-spyware software [4], data backup software [6] and wireless security settings [37] have used a decomposed measure of risk to predict adoption intention. That is, based on the Protection Motivation Theory (PMT) [27], these studies measure the perceived vulnerability (likelihood) and severity (impact) of an adverse event, rather than the perceived risk. The same approach is also found in research on antecedents of employees' security policy compliance [28,11,2]. Although the measures used are more specific on security threats and also include adversaries, they do not include perceptions of current security measures or adversaries. The foundations for risk perception in psychology is further discussed in the upcoming section.

There is an abundance of risk analysis frameworks that describe in varying degree of detail the considerations that should be taken when analysing risks. Examples of such frameworks include the OWASP Testing guide [24], OCTAVE Allegro [3], NIST RMF [36] and CORAS [20]. The focus of these frameworks is to guide corporations or security professionals in performing thorough, comprehensive and reliable risk analyses. The aim is therefore to capture a more objective assessment of security risks. For the purpose measuring subjective risk perception, these frameworks cannot be applied directly. However, they do provide valuable insights into the concept of risk and provide an important basis on which we build our risk perception instrument.

Our work differ from previous research in that it attempts to bridge the gap between risk perception measures and risk analysis methodologies. The risk perception measure we propose maintains the subjective notion of risk while being more concrete and more detailed concerning the risk concept itself. Thus we extend risk perception measures to include the common concepts of risk found in formal risk analysis methodologies.

3 Background

Since risk is all about handling uncertainties, we provide a brief introduction to the psychological process of judgement under uncertainty. The cognitive processes form a basis for understanding the existing theories of risk perception. Finally, we give an overview of the main common concepts of risk found in existing risk analysis methodologies. This is by no means intended as a complete review of either area, but rather to identify some important concepts that we use when devising a risk perception measure.

3.1 The Psychology of Judgment under Uncertainty

There seem to be general consensus among researchers that the cognitive process may be regarded as two separate, although connected, systems [16]. One is the fast and effortless *intuition* while the other is the slow and time consuming *reason*. Epstein [7] have called these systems the *experiental system* and the *rational system*, whereas others including Kahneman and colleagues [17,16] refer to them simply as *system 1* and *system 2*. These two systems are assumed to operate

quite differently. The intuition creates the initial judgment, whereas the rational system monitors and correct obvious mistakes. Intuition is fast, automatic, effortless, emotional and capable of parallel operations, whereas the rational system is slow, effortful, conscious and controlled. Thus, intuitive judgments can be made in parallel with other tasks, without giving it much conscious thought, while rational cognition requires more attention and therefore often interrupts other cognitive tasks. That is seemingly why the intuitive system always attempts to make a suggestion, while the rational system monitors and corrects the judgments that are obviously wrong. In general Kahneman reports five ways in which a judgment is made [16]:

1. An intuitive judgment or intention is initiated, and
   (a) Endorsed by the rational system;
   (b) Adjusted (insufficiently) for other features that are recognized as relevant;
   (c) Corrected (sometimes overcorrected) for an explicitly recognized bias; or
   (d) Identified as violating a subjectively valid rule and blocked from overt expression.
2. No intuitive response comes to mind, and the judgment is computed by the rational system.

Whenever our cognition fails, it is therefore not only the failure of one system, but both. First, the intuition fails to give proper judgment, either by giving an erroneous one or by not giving one at all. Next, the rational system fails to correct the initial intuitive judgment or fails to compute a proper judgment if no initial intuitive response was created.

3.2 Rational Risk Perception

Fischhoff, Slovic et al. [10,31,35,30] have studied antecedents or dimensions of risk perception and their influence on the perception of risk and benefit, and found that these dimensions were highly correlated. For instance, it seemed that risks that where perceived to be controllable also was perceived voluntary. Based on the correlation between dimensions, Fischoff and Slovic proposed two sets of dimensions, or factors, affecting risk perception. This has later been termed the *psychometric approach* and have been one of the key approaches to understanding risk perception. High risk is characterized by factor 1 as events which are unobservable and new, have unknown risk and delayed effect. On factor 2 high risk is characterized as events that are uncontrollable, involuntary and difficult to prevent, and have catastrophic fatal consequences involving considerable dread.

The *psychometric approach* is based to a large extent on the idea that risk perception is a conscious deliberate rational process. That being said, *dread*, being intimately associated with intuition, was found to be one of the dominant dimensions of this approach. Still, as explained by Slovic later the focus of the research was on rational cognition [32].

Critics of the approach [29] have pointed to a number of weak points, particularly the fact that factor analysis is done on average ratings of dimensions,

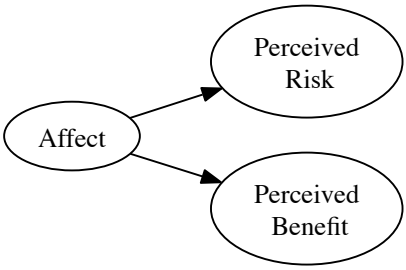


Fig. 1. The affect heuristic [33,34]

instead of individual ratings. Although this does not necessarily mean that the factors identified in the psychometric approach are wrong, it implies that there could be many more factors than what originally was found.

3.3 Risk and Affect

From the theory of decisions under uncertainty it is apparent that intuition also plays a vital part in understanding risk perception. Affect was identified as part of the *psychometric approach* (e.g. dread and fear), but has later thought to be more important in determining risk perceptions. Alhakami and Slovic (reported in [34]) found that although risk and benefit often are positively correlated in real life (i.e. great risk comes with great benefits), people's perception of risks and benefits are often negatively correlated. This discovery led them to propose that risk and benefit perceptions were guided by the affective impression of activity or technology in question. Later, this effect was termed the *affect heuristic* for risk perception [9,33,34]. The basic idea of this heuristic is that people record representations of prior events that are labeled according to the affective response they generate. Thus, whenever a judgment is to be made, people consult the recordings and are guided by the positive or negative affective label attached to it [9]. Positive feelings yield high perceived benefit and low perceived risk (see Fig. 1).

Finucane et al. [9] provided empirical evidence for the affect heuristic by giving different information concerning the risks and benefits of nuclear power. The study showed that whenever participants were informed of the benefits, they perceived the risks to be lower than when not receiving such information. In the same paper, the authors also used time limitations to demonstrate the affect heuristic, since system 1 operates much faster than system 2, the study suggests that under time pressure, people will rely on affect for their risk judgments.

Loewenstein et al. [19] propose the *"risk as feelings"*-hypothesis that feelings and cognitive evaluations are mutually influenced by each other and they both have a direct effect on behaviour, positing that emotions often produce behaviour that is in conflict with optimal behaviour. Fig. 2 illustrates these close interconnections of the two systems on risky behaviour.

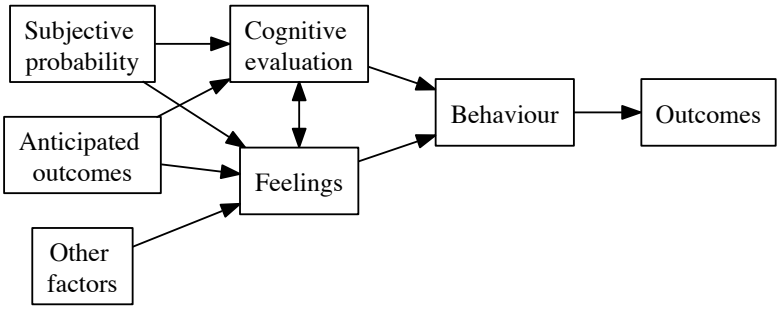


Fig. 2. Risk as feelings model [19]

3.4 Risk Analysis

As mentioned in Section 2 there are several existing frameworks for conducting risk analyses. Although they are generally unsuitable to be directly transferred to a risk perception measure, the fundamental concepts on which they rely are useful. We therefore outline the predominant definitions of risk and discuss the contents of the main concepts of information security risk.

Definitions. Although there are many definitions of risk in the area of information security, they all share the basic concepts of likelihood and impact. For example, ISO defines the *level of risk* as the *"magnitude of a risk, expressed in terms of the combination of consequences and their likelihood"* [13], whereas the the Open Web Application Security Project (OWASP) denotes this product of likelihood and impact as *risk*. Similarly, NIST defines risk as *"the net mission impact considering (1) the probability that a particular threat-source will exercise (accidentally trigger or intentionally exploit) a particular information system vulnerability and (2) the resulting impact if this should occur"*[36]. Despite the different wording the general concept that risk concerns the combined likelihood and impact of an adverse event is generally agreed upon (see Fig. 3).

Impact. The impact constitutes the negative effects an adverse event would have. The OWASP testing guide [24] distinguishes two separate kinds of impacts when determining the overall risk. Technical impact includes a breach of one or more of the security attributes such as loss of confidentiality, integrity or availability. Business impact on the other hand, is concerned with the resulting specific damage to the business or organisation. ISO-27005 [13] and NIST 800-30 [36] refer to the impact as the *business impact* and use *breach of information security (goals)* for the *technical impact*. The meaning however remains largely the same.

The technical impact is largely unaffected by the context or domain in question. Data theft will always yield loss of confidentiality (or we would not call

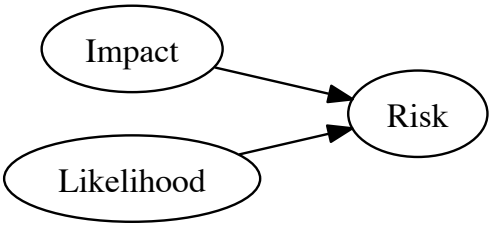


Fig. 3. The basic risk model

Table 1. Technical and business impact

Technical impact	**Business impact**
Breach of security attribute	Breach of business goals
Domain and context insensitive	Domain and context sensitive
Security knowledge	Business/domain knowledge

it theft) as the technical impact, but the impact this will have on the business is very much dependent on the domain, organisation and context in which the adverse event occurred. Since the technical impact of adverse events remains relatively stable across organisations and domains it is difficult to dispute it, whereas the impact on the business may be much less obviously identified and hence often also disputed and debated. Finally, it requires security knowledge to identify the technical impact, whereas business knowledge is required to identify the corresponding business impact. Table 1 gives an overview of these differences.

There are several factors that could be investigated in order to assess the business impact. OWASP lists four such factors; financial damage, reputation damage, non-compliance and privacy violations[1]. There could potentially be many more, depending on the domain in question. In a previous study we have shown that loss of competitiveness is seen as the most important business impact of the oil and gas industry [22]. Other potential impact factors include operational discontinuity, legal liability, customer dissatisfaction or harm to personnel.

Likelihood. The other main aspect of risk is the *likelihood* of an adverse event occurring. However, unlike classical safety problems there are few ways of gaining reliable statistical data to accurately calculate the probability. Metrics such as Mean Time To Failure (MTTF) are inherently difficult whenever it involves an actual non-random attacker. The problem is not only that any assessment of attack likelihood is to a great extent dependent on subjective opinions, but that

[1] Although the OWASP testing guide [24] does not explicitly define *financial damage*, we use it here to denote direct monetary losses as a result of an adverse event. This is to prevent confusion with *business impact*.

it may be difficult to even have an opinion. What is the likelihood of foreign intelligence agencies conducting espionage in my company? Unless you have actually experienced it happening, it is very difficult to know what to think.

When performing risk assessments it is thus common to instead look at the factors that affect the likelihood, instead of the actual likelihood itself. The OWASP testing guide [24], ISO27005 [13] and NIST 800-30 [36] provide three sets of factors of risk likelihood:

- *Vulnerability*: Is it easy for an attacker to discover the vulnerability? If it is discovered, is it easy to exploit? Is the vulnerability common knowledge?
- *Attackers or Threat agents*: Are the potential attackers skilled to perform an attack? Do they have a motive? How often does the opportunity to exploit the vulnerability present itself? How many possible attackers are there?
- *Existing controls*: To what extent do the current security measures reduce the likelihood of exploiting the vulnerability? Is it likely that an exploit can be detected?

In the OWASP guide *existing controls* are not explicitly part of the guidance. However, implicitly it may be argued that detection mechanisms may very well be considered a control mechanism.

Some of these factors are highly inter-connected, such as how easy it is to exploit a vulnerability and how skillful the threat agents are. If the exploit is difficult, the threat agents must be very skillful in order to launch an attack; if the exploit is easy, it really does not matter how skillful the threat agents are.

4 Towards a Model and Measure of Risk Perception

Risk can be modelled in many different ways. It is common to first model a construct and next worry how to operationalize and measure it. However, as noted by Jarvis et al. [14] and Petter et al. [26], it is useful to look at these two issues combined rather than in isolation. In this section we define the constructs of a risk perception model that is needed to create a corresponding measurement instrument. In doing so, we discuss the different types of constructs and their implications for the instrument.

4.1 Construct Types

There are fundamentally three different ways in which constructs can be modelled: reflective, formative or multidimensional [14]. The way in which a construct is modelled also affects the way in which the construct can be measured. Below we present the main differences between these different types of constructs.

A *reflective* construct is used to denote a one-dimensional construct (i.e. has no sub-constructs) where the same aspect of the construct can be measured in different ways. In a standard questionnaire-type measure, this would mean that all items or questions measure the same underlying phenomenon and therefore all items are supposed to covary. Therefore, a respondent scoring low on one

scale item is more likely to score low on the other scale items of the construct. Further, the scale items should be replaceable and removing one should not affect the overall measure of the construct. Reflective constructs are the predominant in Information Systems Research [26].

A much less used type of construct is the *formative* construct [26], where the measurement items are combined to *form* the construct. From its definition, a formative construct is defined through its measurement items [14] such that changing one of the items would yield a changed construct. Unlike the reflective constructs, the items of a formative measure are supposed to tap into different aspects of the construct in order to "cover" the entire construct. Hence, the measures are not required to and often not intended to covary.

Multi-dimensional constructs are modeled as a composite of its sub-constructs where the construct is defined through its sub-constructs. Each sub-construct may in turn be modeled as either reflective, formative or multidimensional. Thus, measuring a multidimensional construct means to combine the measures of all its sub-constructs, representing a kind of divide-and-conquer strategy to modelling.

4.2 A Risk Perception Model

The model of risk perception we propose is based on the concepts identified in common risk analysis frameworks and previous research on risk perception. The model is depicted in Fig. 4. Below, we describe the different constructs of the model and suggest which type of construct would be appropriate. Determining the correct type of construct depends on the purpose of the model and of course how it is intended to be measured. Thus, it may be perfectly sensible to model the constructs of our model differently, if the purpose was different.

Risk is in current literature and among security professionals typically defined as the combination of likelihood and impact (see Section 3). If we were to model risk as reflective, the corresponding measure would have to be on the form *"How risky is it to ... ?"*. There would be no way to capture the likelihood and impact of a threat. Thus, for our purposes, it is evident that risk should either be formative or multi-dimensional. The difference between these two types are more subtle, since it really depends on whether affect, likelihood and impact are just three measures that jointly form risk perception, or if they are seperate constructs to be measured on their own. As we base ourselves on the concepts of risk analyses, we believe the latter to be the case and hence model risk perception as a multidimensional construct.

The concept of *affect* is commonly viewed in risk perception literature as the *"experienced feeling states associated with positive or negative qualities of stimulus"* [32, p. 4]. Although seldom explicitly labeled as such, it is typically assumed to be reflective. That is, measuring feelings such as fear or worry all tap into the same aspects of a concept (the negative qualities).

Likelihood is similar to risk perception in that there are different concepts that make up the likelihood (Fig. 4). However, a notable difference is that likelihood is not commonly defined as the combination of the vulnerabilities, attackers and existing controls. These three concepts influence the likelihood, although

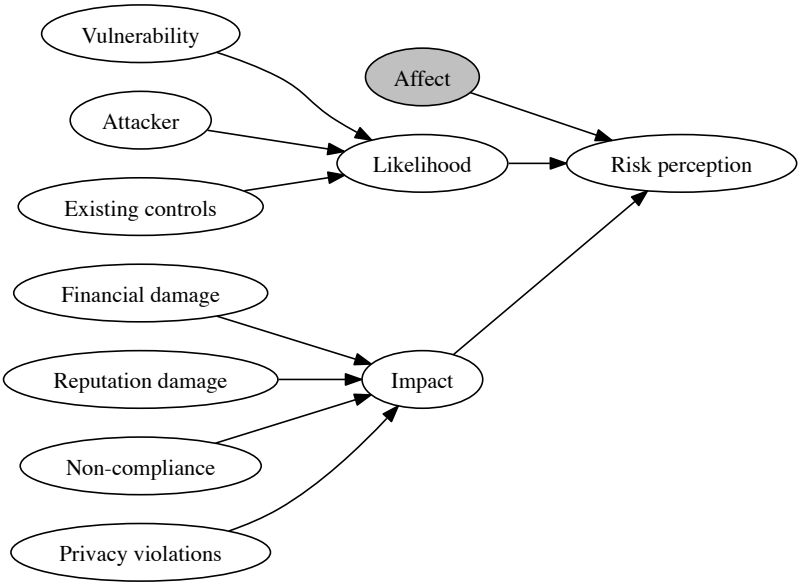


Fig. 4. Concepts of information security risk

not necessarily simultaneously. I.e., it is perfectly conceivable that there may exist serious vulnerabilities and no motivated or skilled attackers at the same time. Thus, the vulnerabilities, attackers and existing measures are not required to covary and hence a formative or multi-dimensional construct seems appropriate. For multi-dimensional constructs it is assumed that the sub-constructs themselves are distinct [14], so that measurements of each sub-construct do not capture each other. In other words, there should be a clear and definite separation between the sub-constructs. Vulnerabilities and existing security measures do however have commonalities even though the concepts themselves are not identical. E.g., the lack of a security measure may increase the vulnerability. Similarly, vulnerabilities and attacker motivation/skill also share some of the ideas. A vulnerability that is easy to exploit would not require the attacker to be very skillful, and hence the two concepts also tap into one another. Therefore, it appears that modeling likelihood as a formative construct may be the better option, such that vulnerabilities, attackers and existing controls denote different aspects of likelihood, rather than complete constructs of their own.

Impact is perhaps the most difficult concept to model, because it depends to a large extent on the threat in question and the context in which it exists. If we are considering the threat of credit card fraud against individual users, it safe to assume that reputation damage is not relevant. However, if we are considering the same threat from the perspective of a payment company, then the reputation damage may be very relevant. The possible impacts (financial damage, reputation damage, non-compliance and privacy violation) in our model are by

no means exhaustive, there could be several other impacts that are important for the context at hand. We have previously shown [22] that users collaborating across company borders are focused on the loss of competitive advantage impact. Thus, the aspects of impact that we list in Fig. 4 are to be treated as a starting point on which context-dependent impact aspects may be added. Since there is no reason to assume that the different possible impacts covary, it does not seem appropriate to model the construct as reflective. Further, it is also noted that financial damage and reputation damage may overlap, in the sense that damage to ones reputation yields reduced sales and thus financial loss. Therefore, we suggest that the impact construct be regarded as a formative construct.

4.3 An Instrument for Measuring Risk Perception

The motivation for modelling risk perception is to devise a measurement framework that can more accurately capture the risk perception of regular users. For this, we propose a self-report questionnaire to let users respond to statements regarding the risk of specific ICT threats. Here we present a preliminary sketch of the instrument, currently with excessive redundancy in the questions. Readers should take care not to treat any of these questions as tested and verified.

In Table 2 we have listed some example questionnaire items for measuring two of the sub-constructs of risk perception; likelihood and impact. The statements are supposed responded to by indicating the degree to which respondents agree with the statements (e.g.on a five-point scale from *"strongly agree"* to *strongly disagree*). Since risk perception is modeled as a multi-dimensional construct, there are no measurement items for the risk perception itself. Instead, it is to

Table 2. Example questionnaire items for the multi-dimensional risk perception construct

Construct	Concept	Statement/Question
Likelihood	Vulnerability	*"Exploiting vulnerability is easy"*
		"The vulnerability is well known"
	Attacker	*"Attackers have the necessary skills to exploit the vulnerability"*
		"Attackers have financial interest in exploiting the vulnerability"
	Existing controls	*"Existing security measures prevent attackers from exploiting the vulnerability"*
		"Existing security measures can detect attackers exploiting the vulnerability"
Impact	Financial damage	*"Attackers exploiting vulnerability would result in financial damage"*
	Reputation damage	*"Attackers exploiting the vulnerability would damage the company's reputation"*
	Privacy violation	*"Attackers exploiting the vulnerability would cause a breach of customer's privacy"*

be computed from its sub-constructs, which in turn are measured through their supporting concepts rather than directly. Hence, we have not included questions or statements directly targeting likelihood for instance. The example questionnaire items do not include measures of affect, as there are several tested and validated measurement instruments that can be used (e.g., Crites et al. [5]).

What becomes apparent when creating measurement instruments is that the threat and potential impact must be specified to a certain detail. It is difficult to say anything about attacker motivation, unless you know who the attacker is. Thus, for the risk perception measure to work, there needs to be a serious effort up front to determine the possible attackers, vulnerabilities, threats and the possible resulting impact. However, by expanding and extending this framework of risk perception measures, the effort required for adapting measurement instruments will steadily decline.

5 Discussion

In this section we discuss some of the difficult issues of validity and trust.

5.1 Construct Validity

A central aspect to any model and measurement is their validity. Does the model actually represent risk perception? Does the measurement actually measure risk perception? Both the model and the measures are based on extensively used and agreed-upon concepts for conducting risk analysis. Our risk perception model is therefore tightly coupled with the information security communities definition of risk. The central point here is whether the concepts of risk analysis can meaningfully be applied to risk perception. Or to put it in other terms: Do people actually consider attackers' motivation as part of their risk perception? In our experience regular users often refer to possible attacker motivations or skills. Statements like *"who would want my information"* or *"it is very difficult to make use of our source code"* are indeed not uncommon and clearly demonstrate that users consciously consider the motivation and skills of potential attackers.

Another thing to consider is that the systematic approach of risk analysis attempts to identify objective risk, that is free from subjective judgments. That is, risk analysis is supposed to trigger the rational system, not the intuitive system. Whereas our risk perception measure also consider affect or intuition. A central concept in the *risk as feelings* theory (see Section 3) is the mutual relationship between the cognitive and affective response to risk. That is, that the rational system both influence and is influenced by the intuitive system.

Our model and corresponding instrument for measuring risk therefore have a sound basis in theory. It is worth noting that the measurement instrument needs to be tested and properly evaluated before definitive statements can be made regarding the validity of the constructs.

5.2 Content Validity

A potential problem with our framework is that it requires identification of the specific threats, vulnerabilities, attackers and potential damages to be assessed. We have already discussed the implications of this in terms of reusing measurement instruments and the effort required to identify and prioritize the concepts. However, a perhaps more severe problem is the potential threat to validity it poses. That is, how do we know that the identified threats, vulnerabilities, attackers and potential damages are appropriate and sufficient? It may be that an important group of attackers are omitted or disregarded from the measurement instrument, which again may give an invalid result. This may happen even if the the statements (or scale items) used have previously been tested and verified. This further strengthens our belief that a risk perception measure should be treated as a starting point and would require great care to address the validity concerns when applied to other contexts.

5.3 Convergent and Discriminant Validity

Our model, constructs and corresponding measures are supported by existing theory and practice. However, it is common to test a new measure for convergent and discriminant validity. That is, to ensure that our measure of risk correlates with other measures that it theoretically should correlate with, while at the same time be unrelated to other measures that theoretically should be unrelated.

However, there are no well established measures of risk perception other than the pure reflective measure risk commonly used (see related work in Section 2). We have postulated that these measures are not particularly good for ICT risk perception as they are too abstract to meaningful for our purposes. Normally one would have liked our measure to have high convergent validity with these other risk perception measures. However, this may indicate that our measure is equally poor as the existing ones. On the other hand, if our measure and the other measures demonstrate discriminant validity (they are unrelated) that could would indicate that one of the measures are invalid, but not which one.

Another method to indicate such validity would be to measure risk perception in relation to another construct. The protection motivation theory for instance postulates that there is a relation between the perceived risk and the intention to use protective technology. Hence, if our measure fits such a model, it would strengthen the argument that the measure is indeed valid.

5.4 Relation to Trust

Risk perception is to some extent related to trustworthiness and trust. There are several trust models that incorporate risk [15], and intuitively trust also imply a certain degree of risk. However, as noted by Mayer, Davis and Schoorman [21], risk is an outcome of trust and therefore "differentiates the outcomes of trust from general risk-taking behaviors because it can occur only in the context of a specific, identifiable relationship with another party" [21]. Since we propose

a general purpose adaptable instrument for risk perception measurement, we have omitted trustworthiness as part of our risk perception. However, in specific contexts where there are indeed identifiable relationships between parties, trustworthiness may well be included as a factor that affect risk perception.

6 Conclusion

In this paper we have presented a risk perception model and a preliminary measure of the risk perception of ordinary ICT users. Our approach is based on a combination of *prior research* on general risk perception and *common practices* in the risk analysis field. The risk perception measure focus on measuring aspects on which ordinary users can be expected to have an opinion. The validity of the construct is promising since it is based on solid existing theory and practice. However, great care should be taken to address the threats to content validity for different risks.

In further work we will employ our instrument in a larger study, testing the validity of our risk perception measure in a larger population.

References

1. Bélanger, F., Carter, L.: Trust and risk in e-government adoption. The Journal of Strategic Information Systems 17(2), 165–176 (2008)
2. Bulgurcu, B., Cavusoglu, H., Benbasat, I.: Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness. MIS Quarterly 34(3), 523–548 (2010)
3. Caralli, R.A., Stevens, J.F., Young, L.R., Wilson, W.R.: Introducing octave allegro: Improving the information security risk assessment process. Technical report CMU/SEI-2007-TR-012, Software Engineering Institute, Carnegie Mellon University (May 2007)
4. Chenoweth, T., Minch, R., Gattiker, T.: Application of protection motivation theory to adoption of protective technologies. In: 42nd Hawaii International Conference on System Sciences, HICSS 2009, pp. 1–10 (January 2009)
5. Crites, S.L., Fabrigar, L.R., Petty, R.E.: Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues. Personality and Social Psychology Bulletin 20(6), 619–634 (1994)
6. Crossler, R.: Protection motivation theory: Understanding determinants to backing up personal data. In: 2010 43rd Hawaii International Conference on System Sciences (HICSS), pp. 1–10 (January 2010)
7. Epstein, S.: Integration of the cognitive and the psychodynamic unconscious. American Psychologist 49(8), 709–724 (1994)
8. Featherman, M.S., Pavlou, P.A.: Predicting e-services adoption: a perceived risk facets perspective. International Journal of Human-Computer Studies 59(4), 451–474 (2003)
9. Finucane, M.L., Alhakami, A., Slovic, P., Johnson, S.M.: The affect heuristic in judgements of risks and benefits. Journal of Behavioral Decision Making 13(1), 1–17 (2000)

10. Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., Combs, B.: How safe is safe enough? a psychometric study of attitudes towards technological risks and benefits. Policy Sciences 9, 127–152 (1978), doi:10.1007/BF00143739
11. Herath, T., Rao, H.R.: Protection motivation and deterrence: a framework for security policy compliance in organisations. European Journal of Information Systems 18(2), 106–125 (2009)
12. Horst, M., Kuttschreuter, M., Gutteling, J.M.: Perceived usefulness, personal experiences, risk perception and trust as determinants of adoption of e-government services in the netherlands. Computers in Human Behavior 23(4), 1838–1852 (2007)
13. ISO/IEC 27005: Information technology — Security techniques — Information security risk management. International Organisation for Standardisation, Geneva, Switzerland (2011)
14. Jarvis, C.B., MacKenzie, S.B., Podsakoff, P.M.: A critical review of construct indicators and measurement model misspecification in marketing and consumer research. Journal of Consumer Research 30(2), 199–218 (2003)
15. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision Support Systems 43(2), 618–644 (2007)
16. Kahneman, D.: A perspective on judgment and choice: Mapping bounded rationality. American Psychologist 58(9), 697–720 (2003)
17. Kahneman, D., Frederick, S.: Representativeness revisited: Attribute substitution in intuitive judgment. In: Gilovich, T., Griffin, D., Kahneman, D. (eds.) Heuristics and Biases: The Psychology of Intuitive Judgment, pp. 49–81. Cambridge University Press (2002)
18. Kim, D.J., Ferrin, D.L., Rao, H.R.: A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decision Support Systems 44(2), 544–564 (2008)
19. Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N.: Risk as feelings. Psychological Bulletin 127(2), 267–286 (2001)
20. Lund, M.S., Solhaug, B., Stølen, K.: Model-Driven Risk Analysis: The CORAS Approach. Springer (2011)
21. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. The Academy of Management Review 20(3), 709–734 (1995)
22. Nyre, Å.A., Jaatun, M.G.: Usage control in inter-organisational collaborative environments – A case study from an industry perspective. In: Quirchmayr, G., Basl, J., You, I., Xu, L., Weippl, E. (eds.) CD-ARES 2012. LNCS, vol. 7465, pp. 317–331. Springer, Heidelberg (2012)
23. Nyre, Å.A., Jaatun, M.G.: On the adoption of usage control technology in collaborative environments. In: Proceedings of the 12th International Conference on Innovative Internet Community Systems, Trondheim, Norway, June 13-15, pp. 142–153 (2012)
24. OWASP: OWASP testing guide v3. Tech. rep., The Open Web Application Security Project (2008), `https://www.owasp.org/images/5/56/OWASP_Testing_Guide_v3.pdf`
25. Pavlou, P.A.: Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. International Journal of Electronic Commerce 7(3), 101–134 (2003)
26. Petter, S., Straub, D., Rai, A.: Specifying formative constructs in information systems research. MIS Q 31(4), 623–656 (2007)
27. Rogers, R.W.: A protection motivation theory of fear appeals and attitude. Journal of Psychology 91(1) (1975)

28. Siponen, M., Pahnila, S., Mahmood, A.: Employees' adherence to information security policies: An empirical study. In: Eloff, M., Labuschagne, L., Eloff, J., von Solms, R. (eds.) Fundamentals of Artificial Intelligence. IFIP, vol. 232, pp. 133–144. Springer, Heidelberg (1986)
29. Sjöberg, L., Moen, B.E., Rundmo, T.: Explaining risk perception: An evaluation of the psychometric paradigm in risk perception research. Rotunde, vol. 84. Norwegian Univeristy of Science and Technology (2004)
30. Slovic, P.: Perception of risk. Science 236(4799), 280–285 (1987)
31. Slovic, P., Fischhoff, B., Lichtenstein, S.: Rating the risks. Environment 21(3), 14–20, 36–39 (1979)
32. Slovic, P.: The Feeling of Risk - New perspectives on risk perception. Earthscan, London (2010)
33. Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G.: Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. Risk Analysis 24(2), 311–322 (2004)
34. Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G.: The affect heuristic. European Journal of Operational Research 177(3), 1333–1352 (2007)
35. Slovic, P., Fischhoff, B., Lichtenstein, S.: Why study risk perception? Risk Analysis 2(2), 83–93 (1982)
36. Stoneburner, G., Goguen, A., Feringa, A.: Risk management guide for information technology systems. NIST Special Publication 800-30, National Institute of Standards and Technology (2002)
37. Woon, I., Tan, G., Low, R.: A protection motivation theory approach to home wireless security. In: Proceedings of the Twenty-Sixth International Conference on Information Systems, pp. 367–380 (2005)

A Framework for Combining Problem Frames and Goal Models to Support Context Analysis during Requirements Engineering

Nazila Gol Mohammadi, Azadeh Alebrahim, Thorsten Weyer, Maritta Heisel, and Klaus Pohl

Paluno - The Ruhr Institute for Software Technology, University of Duisburg-Essen, Germany
{nazila.golmohammadi,azadeh.alebrahim,thorsten.weyer, maritta.heisel,klaus.pohl}@paluno.uni-due.de

Abstract. Quality requirements, like security requirements, are difficult to elicit, especially if they cross multiple domains. Understanding these domains is an important issue in the requirements engineering process for the corresponding systems. Well-known requirements engineering approaches, such as goal-oriented techniques provide a good starting point in capturing security requirements in the form of soft-goals in the early stage of the software engineering process. However, such approaches are not sufficient for context and problem analysis. On the other hand, the context and problem modeling approaches like e.g., problem frames, do not address the system goals. Integrating the relevant context knowledge into goal models is a promising approach to address the mutual limitations. In this paper, we propose a framework for combining goal models and problem frames. The framework makes it possible to document the goals of the system together with the corresponding knowledge of the system's context. Furthermore, it supports the process of refining (soft-) goals right up to the elicitation of corresponding security requirements. To show the applicability of our approach, we illustrate its application on a real-life case study concerning Smart Grids.

Keywords: Requirements engineering, security requirements, problem frames, goal modeling.

1 Introduction

Requirements engineering (RE) is an engineering activity that ties up the development activities with the real-world problems. It performs a series of activities based on the recognition of a problem to be solved and leads to a detailed specification of that problem [1]. This specification includes the requirements of the system-to-be in a way that the system meets the expected quality concerns of the stakeholders as well as the functional ones. Although the treatment of quality requirements in software development is not yet as well mastered as the treatment of functional requirements [2], it has recently caught more attention. Quality requirements such as security requirements must be elicited, analyzed, and documented as thoroughly as functional ones.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 272–288, 2013.

Understanding the context is an important issue in RE for eliciting and analyzing requirements. During context analysis, the analyst anticipates how the system-to-be will be integrated in its real world context, once it is in operation. Due to the fact that the size and complexity of the environment (context) of typical software-intensive systems, e.g., as integrated in smart grid, grew very fast in the last decade, e.g., systems-of-systems, the context analysis activities during RE become more and more crucial. Context analysis techniques support the requirements engineer to specify the requirements of the system in a systematic way. Therefore, the first mandatory task in RE is to know and explore the context, which the system will interact with. Focusing on the system itself prior to modeling its context lacks considering domain knowledge in eliciting requirements. The elicited requirements might be built on incomplete or even wrong domain knowledge. This can lead to failure of the system due to incorrect or incomplete requirements. Considering security requirements, this might be even more costly and disastrous because of business critical data losses and violation of security rules or any unacceptable outcome.

Goal modeling approaches support the elicitation of system purposes in the RE process, but they have shortcomings as well. First, they tend to be inadequate and imperfect in describing the problem context [3]. Second, the goal models grow quickly, posing a threat to manageability and traceability [4]. Third, goals are hierarchical, hence, it sometimes becomes difficult to determine where a goal is situated in the hierarchy and how it relates to the problem context. Moreover, for every goal, there is always a discoverable super-goal. Thus, goal modeling techniques need to be bounded by the problem domain [5].

Problem frames [6] are concerned with the context modeling and the understanding of the context in RE. They analyze the problem context as it is in the real world. However, they lack addressing the goals of the system under consideration.

Therefore, problem frames have the potential to be combined with other RE techniques, such as goal-oriented approaches [7,8]. Goal-oriented approaches investigate goals of the system, whereas in problem frames goals can be identified by investigating the context of the problem. The relationship and the gap we identified between problem frames and goal-oriented approaches suggest the combination of these techniques. In the problem frames approach, there is no way to connect the requirements to the goals or to identify new goals. The goals can be identified by determining why we want to find a solution for this problem. After finding the way to identifying the goal, which is relevant to analyze the specific problem, the goal could be connected to the requirements.

On the other hand, both approaches are not able to show the dynamic behavior of the situation. They are not expressive in showing requirements behavior and the real world interactions. In this paper, we propose a framework based on problem frames and goal models with support of Message Sequence Charts (MSCs) [9] that provides support for the RE phase. The dynamic behavior of the problem context is modeled using MSCs. The gaps and drawbacks mentioned above are addressed using our framework. This combination can provide a good

understanding and overview of the real world prior to structuring the system and map the problem to its solution.

The remainder of this paper is organized as follows: In Section 2 we explain the fundamentals of our method. Section 3 presents our approach for combining problem frames and goal models to support context analysis during requirements engineering. To demonstrate the usefulness of the framework we demonstrate its application on a case study taken from the smart grid domain in Section 4. Section 5 discusses related work. Finally, Section 6 gives a conclusion and sketches future work.

2 Fundamentals

In this section we briefly introduce the problem frames approach, goal modeling, and the use of MSCs in RE. These techniques are the fundamentals for the framework that is described in Section 3.

2.1 Problem Frames

Problem frames [6] are one of the RE approaches, located in the early life-cycle of software engineering. Problem frames capture information from the context of the system in a simple way. A problem frame consists of *domains*, *interfaces* between them, and a *requirement*. Domains describe entities in the real world that are distinguished in three different types: *biddable domains* that are usually people, *causal domains* that comply with some physical laws, and *lexical domains* that are data representations. *Interfaces* connect domains, and they contain *shared phenomena*. Shared phenomena may be events, operation calls or messages. They are observable by at least two domains, but controlled by only one domain, as indicated by the name of that domain and "!". An example can be find in Section 4 (see Fig. 6).

When we state a requirement, we want to change something in the world with the *machine* (e.g., software) to be developed. Therefore, each requirement constrains at least one domain. Such a constrained domain is the core of any problem description because it has to be controlled according to the requirements. Hence, a constrained domain triggers the need for developing a new software (the machine) that provides the desired control. A requirement may refer to several domains in the environment of the machine. The task is to construct a machine that improves the behavior of the real world (in which it is integrated) in accordance with the requirements. According to the problem frames approach, the problem is decomposed into subproblems, which are represented by *problem diagrams*. A problem diagram consists of a submachine, the relevant domains, the interfaces between these domains, and a requirement.

In this paper we describe problem frames using UML4PF (UML profile for Problem Frames) [10,11] that is based on UML class diagrams. In Section 4, we model the context of our application example using the UML notation for problem frames.

2.2 Goal Modeling

In the RE community, goal modeling approaches have gained considerable attention. These approaches aim at capturing the rationale for the software system development. Goal models are mostly represented in tree-like structures that define the intentions of different stakeholders at different levels of abstraction [12]. Goal modeling links the high-level goals to the low-level requirements [13]. Goal-oriented approaches, such as KAOS [13], i* [14] and goal graphs [12], seek reasons or purposes to design the system (purpose of the system). These approaches use AND/OR connectors to represent the goal decomposition and to define the alternative solutions for fulfilling the super-goal. Goals can be classified into two different categories: hard-goals and soft-goals. Hard-goals may refer to the functional properties of the system behavior, whereas, soft-goals represent quality preferences of the stakeholders. Soft-goals can be achieved at different levels of satisfaction, which means that there is no clear-cut definition for their satisfaction. Examples can be find in Section 4 (see Fig. 5 and Fig. 8). In KAOS notation the goals are modeled as parallelograms and soft-goals in form of dashed parallelograms. Among different goals in the hierarchy of KAOS goal model, there are either AND-refinement or OR-refinement. A positive or negative influence on a soft-goal can be represented using +, ++ or -, - -. A contribution operator represents a positive (+, ++) or negative (-, - -) influence on soft-goals. It documents to which level of satisfaction a soft-goal is influenced either positively or negatively.

2.3 Message Sequence Charts

MSCs [9] have proven to be successful in system development, in particular for RE. They are widely used for the concrete description of not only the system behavior, but also the context. They are especially used to describe the context behavior by requirements engineers and later validate the requirements for an acceptance test. In addition, they provide a convenient notation to describe behavioral requirements in the form of interaction diagrams. MSC consists of a set of actors/entities of the system (boxes), life-lines (vertical dashed lines), messages between the actors (arrows), and states (ovals). Actors can send and receive messages ordered by their occurrence within their life-lines. The state of different actors may be changed upon to receiving a message. An example can be found in Section 4 (see Fig. 7). When using MSCs for modeling behavioral requirements the corresponding diagram specifies the flow of interactions between the system and its context that has to be performed to satisfy a specific purpose of the system.

3 Framework for Combining Problem Frames and Goal Models to Support Context Analysis

We show how the problem frames approach can be enhanced by linking goals to the problem diagrams satisfying that goal. The goal in the KAOS goal model

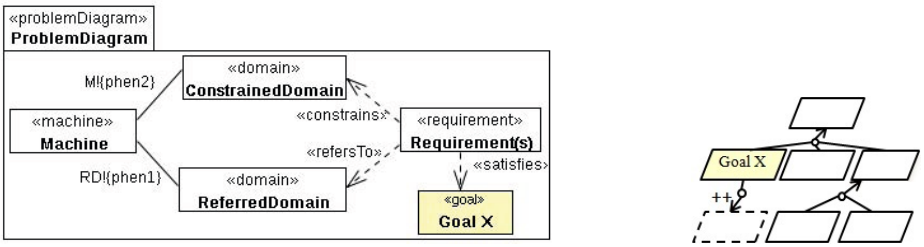


Fig. 1. Relation between the goal and related requirement(s)

(Goal X in Fig. 1) will be localized by setting up a problem diagram. By using the established problem diagram, domains that contribute to the satisfaction of Goal X will be identified. The Goal X in the KAOS model is modeled in the problem diagram, as well as using the stereotype ≪`goal`≫. The *Requirement(s)* in the problem diagram (Fig. 1) is a placeholder for requirements, which will be elicited when using our method. Requirement(s) for satisfying the goal under consideration (e.g., Goal X) will be elicited with the knowledge gained from domain analysis based on domains and phenomena between them. The use of MSC models helps to represent the dynamics of the problem diagrams. The goal model that will be further refined is based on the knowledge gained from the domain analysis using the problem diagram. Our framework provides support for the requirement elicitation considering its respective goal. Note that this process is iterative and recursive. It means that the new refined goals (added to the goal model) will be selected in following iterations for deriving refined requirements. In Section 3.1, our method is described in a detailed and methodic way. In Section 3.2, we show where our method is located in the RE technology map proposed by Nakatani et al. and Tsumaki et al. [15,16].

3.1 Combining Problem Frames and Goal Models to Support Context Analysis

- **Step 1 — Set up the Goal Model:** The first step is concerned with setting up the goal model. We model and document the goal model using KAOS. This model is established based on stakeholder intentions, which include the purpose of the system under development. The goals are captured either by interviewing involved stakeholders or based on expertise of a requirements engineer. We start with high-level goals, and then refine them into hierarchical goal structures.
- **Step 2 — Select a Goal:** After setting up the KAOS goal model, we select a goal (e.g., Goal X). This goal can be one of the major concerns of stakeholder(s), which should be followed systematically until its satisfaction is achieved. To this end, the Goal X will be further enhanced with domain knowledge gained in the next steps. Goal X is the input for the next step.

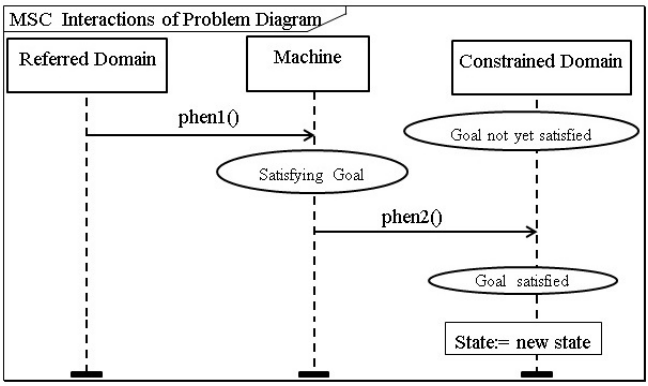


Fig. 2. MSC for the realization of the goal addressed in the problem diagram

- **Step 3 — Identify Relevant Domains:** For the selected goal (Goal X), we identify the domains from the problem context under consideration. These domains are involved in the satisfaction of Goal X. The output of this step is a list of contributor domains that will set up the corresponding problem diagram in the next step.
- **Step 4 — Set up the Problem Diagram:** By using the domains identified in the previous step, we set up problem diagrams. These problem diagrams depict a small and focused portion of the problem context related to the selected goal. This supports a more objective and structured way of addressing the problem rather than having the overall one. In this way, the goal located in the goal model is mapped to at least one problem diagram. In this step, we do not yet know the requirements that satisfy the selected goal. We only keep a placeholder for the requirements that we extract in the next step. Figure 1 on the left-hand side shows how a problem diagram can be linked to the part of the goal model, which is realized using that problem diagram.
- **Step 5 — Extract the Requirements:** By using the domains accommodated in the problem diagram and the related goal, we extract the requirements to achieve the selected goal. The intentions of solving the problem refer to the goal of the system under consideration, which will be satisfied by the requirements in the problem diagram. The extraction is based on the domains and phenomena between them. The requirements are described in natural language. The description of derived requirements include domains and phenomena in the problem diagram. Derived requirements shall specify the expected behaviour of the domains as a reaction to referred or observed phenomena. We then complete the problem diagram using the extracted requirements. The placeholder for requirements in the problem diagram indicates the identifier(s) of extracted requirement(s).
- **Step 6 — Set up the MSC:** In this step, we take the problem diagram annotated with the goal as input. Then, we model and document the

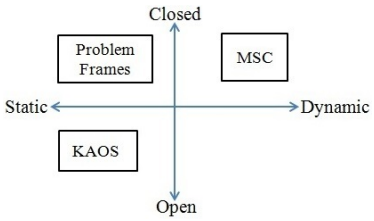


Fig. 3. RE technology map based on [15,16]

interactions of the domains for the realization of the goal. As mentioned in Section 1, the problem diagrams themselves are not able to visualize or represent the interactions between the domains in the required order. The MSCs should capture the following information:

- Initialization. The sequence of actions and events, which would bring the machine to its initial state. In this state, the machine begins its operation. The context domains have also some sets of states.
- The dynamic behavior of the requirement and the interactions between the domains. It is not clear which phenomena will trigger the machine for solving the represented problem.
- Breakage. A causal domain may be damaged if certain sequences of operations are performed on it. Such kinds of sequences should be identified and avoided [17,18].

The MSC related to a problem diagram and its addressed goal shows how the goal can be realized. It includes the context domains identified in the problem diagram. The phenomena in the problem diagram form the messages. The mapping between the problem diagram on the left-hand side of Fig. 1 to the MSC for describing the dynamic behavior and the interaction between the domains in the context is shown in Fig. 2. The actors in the MSC should be found as domains in the problem diagram. The messages between the actors represent the phenomena.

– **Step 7: Refine the Goal Model** In this step, we refine the goal model further, if it is needed. Then, we continue with step two. The refinement of the goal model is shown on the right-hand side of Fig. 1 by dashed lines representing the new identified goal.

3.2 Location of Our Framework in the RE Technology Map

Our approach mainly focuses on the early requirements process dealing with requirements elicitation. Among a number of available RE techniques, we make use of three widely practiced ones: problem frames, goal-oriented, and scenario-based approaches. Our method uses these three techniques that clearly differ from each other in nature. To characterize the differences of these techniques and the benefits of combining them as a framework, we use a map for RE techniques [15,16] shown in Fig. 3. This map has two dimensions. One of these dimensions concerns

the elicitation operation types and the other one concerns target object types. The operational type illustrates how the requirements elicitation and acquisition process is conducted. This is done by focusing on either static structures or dynamic behaviors:

- Static: The static structure of the domains in the context is analyzed for requirements elicitation. The methods belonging to this category, e.g., problem frames and KAOS goal models try to accomplish the RE activities in a structured and systematic way. There are rules or guidelines how to decompose the problem or how to refine the goals, respectively.
- Dynamic: Requirements are elicited from the domain focusing on their dynamic behaviors. Requirements are elicited in an imaginative and unmethodical way like scenarios and interviews. For many stakeholders, it is often easier to think in terms of procedural behaviors or situational changes over time. The dynamic behavior is modeled in our framework using MSCs, which are easy understandable.

The second dimension, the object type, is concerned with the properties of the target context, which has to be analyzed. It can be either closed or open:

- Closed: The target context is relatively stable, known, and closed. This space is basically structured and bounded by a particular syntax. Therefore, it can be understood by focusing on the syntax. Hence, using a closed RE technique bounds the scope of meaning.
- Open: The object space is relatively unstable, unknown, changing, and open. Here, meanings should be well considered for analysis the target context. The approaches belonging to this category are more human oriented, e.g., brainstorming and goal modeling techniques belong to this category.

Figure 3 maps the three techniques used in our approach with regard to the dimensional spaces. Nakatani et al. and Tsumaki et al. [15,16] allocated various RE techniques in this map. The allocation of these techniques in different quadrants in the RE map demonstrates the coverage of our approach that bridges the lower-left quadrant to the upper-right quadrant. We begin in our framework with a KAOS goal model, which is open and static. Then, we carry on with problem frames for decomposing and structuring the context. Problem frames are closed and therefore bound the scope of meaning the KAOS goal models. Subsequently, we use MSCs, which are closed and can express the dynamic behavior.

4 Application Example

In this section we show the application of our approach to the real-life case study "Smart Grids" adapted from the European Network of Excellence on Engineering Secure Future Internet Software Services and Systems (NESSoS)[1].

[1] `http://www.nessos-project.eu/`

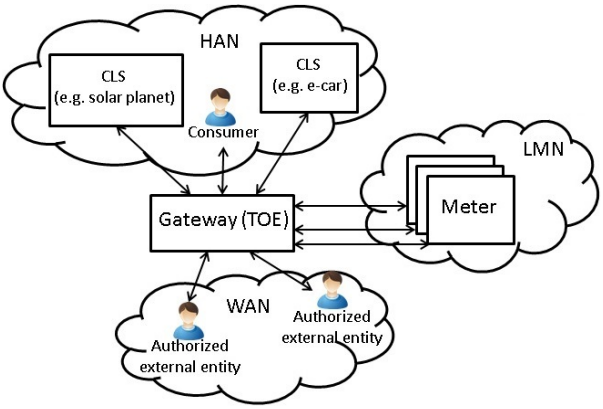


Fig. 4. The context of a smart grid system based on [19]

4.1 Introduction to the Case Study "Smart Grid"

To use energy in an optimal way, smart grids make it possible to couple the generation, distribution, storage, and consumption of energy. Smart grids use Information and Communication Technologies (ICT), which allow for financial, informational, and electrical transactions. In order to define the real functional and security requirements, we considered the documents "Protection Profile for the Gateway of a Smart Metering System"[19] provided by the German Federal Office for Information Security[2] and "Requirements of AMI" [20] provided by the EU project OPEN meter[3]. Figure 4 shows the simplified context of a smart grid system based on [19].

A smart grid involves a wide variety of data that should be treated in a secure way. Protection profile defines security objectives for the central communication unit (Gateway in Fig. 4) in a smart metering system. Beforehand, we define the terms specific to the smart grid domain taken from the protection profile:

Gateway represents the central communication unit in a *smart metering system. It is responsible for collecting, processing, storing, and communicating meter data.*

Meter data *refers to meter readings measured by the meter, regarding consumption or production of a certain commodity.*

Meter *represents the device that measures the consumption and production of a certain commodity and sends it to the gateway.*

Authorized external entity *could be a human or IT unit that communicates with the gateway from outside the gateway boundaries through a WAN. The roles are defined as external entities that interact with the gateway and the meter are consumer, grid operator, supplier, gateway operator, gateway*

[2] http://www.bsi.bund.de
[3] http://www.openmeter.com/

administrator, ... (For the complete list of possible external entities see the protection profile[19]).

WAN (Wide Area Network) *provides the communication network that interconnects the gateway with the outside world.*

LMN (Local Metrological Network) *provides the communication network between the meter and the gateway.*

HAN (Home Area Network) *provides the communication network between the consumer and the gateway.*

Consumer *refers to end user or producer of commodities (electricity, gas, water, or heat).*

4.2 Application of the Approach to the Smart Grid

Applying our approach to the case study implies the following activities and working results.

Step 1 — Set up the Goal Model: The intentions of the stakeholders are reflected in the documents mentioned above. Therefore, they provide a good basis for setting up a goal model for the smart grid system. We set up the goal model by locating the super-goal *Develop Gateway for Smart Meter System* at the top of the goal model. The Gateway has to satisfy the required behavior for such a system. It has to achieve 20 goals as stated in [20]. The goals are divided into three categories *minimum*, *advanced*, and *optional*. Thirteen Goals are contained in the category *minimum* that represents the necessary ones to achieve the super-goal of the system. For reasons of clarity and comprehensibility, we only show three of 20 goals required to achieve the super-goal in Fig. 5.

Step 2 — Select a Goal: We select the goal *Meter Reading for billing*, which is contained in the category *minimum*. According to [20], this goal deals with gathering, processing, and providing of meter reading for the billing process. The selected goal serves as input for the next step.

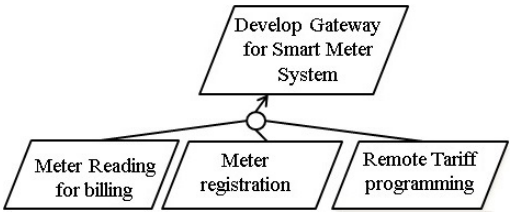


Fig. 5. Simplified goal model for the smart meter system including the super-goal and its directly related goals

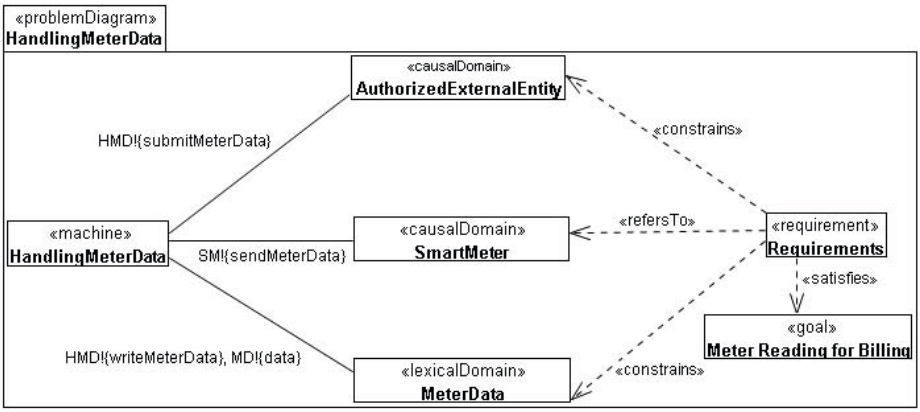


Fig. 6. Linking of the selected goal to the related problem diagram

Step 3 — Identify Relevant Domains: To identify the relevant domains, we make use of the selected goal and the context of the smart grid system. For the gathering of meter readings we need to connect to the *smart meter*. The gathered meter readings have to be processed. Therefore, we need to store them into *meter data*. After processing, the meter data should be provided to the *authorized external entity* in order to create the bill. The domains we identified are therefore *smart meter*, *meter data*, and *authorized external entity*.

Step 4 — Set up the Problem Diagram: We set up a problem diagram, which contains the domains *SmartMeter*, *MeterData*, and *AuthorizedExternalEntity* identified in the previous step. Figure 6 shows the corresponding problem diagram in UML notation. The machine *HandlingMeterData* has to meet the requirement *Requirements*, which leads to satisfying the related goal *Meter Reading for Billing*. The problem diagram describes that the machine *HandlingMeterData* receives meter data from the *SmartMeter*, writes it into *MeterData*, and submits it to the *AuthorizedExternalEntity*. Note that the problem diagram does not make any statements about the order of performing the phenomena. The notation $SM!\{sendMeterData\}$ (between the domains *HandlingMeterData* and *SmartMeter*) means that the phenomenon *sendMeterData* is controlled by the domain *SmartMeter*. The requirement *Requirements* constrains the domains *MeterData* and *AuthorizedExternalEntity*. This is expressed by a dependency with the stereotype ≪`constrains`≫. It refers to the domain *SmartMeter* as expressed by a dependency with the stereotype ≪`refersTo`≫. Note that the requirement *Requirements* acts as a placeholder in this step. It will be extracted in the next step. In this way, the goal located in the goal model is mapped to the context.

Step 5 — Extract the Requirements: Inputs for this step are the domains from the problem diagram and the related goal. The domains *SmartMeter*, *MeterData*, and *AuthorizedExternalEntity* are contained in the problem diagram

HandlingMeterData. The related goal *Meter Reading for billing* is concerned with gathering, processing, and providing of meter reading for the billing process. Based on this information, we identify four new refined requirements, which replace the requirement *Requirements* in Fig. 6.

- R1: The smart meter sends meter data to the Gateway.
- R2: The Gateway shall process the received meter data.
- R3: The Gateway shall store the result of the process into meter data.
- R4: The Gateway shall submit the result to external parties.

Step 6 — Set up the MSC: In this step, we model and document the dynamic behavior of the context using MSCs. We represent the sequence of phenomena in the initialization of the problem and the realization of the goal. Figure 7 shows the dynamics of the goal *Meter Reading for billing*. The same domains depicted in the problem diagram become actors in MSC and related phenomena are the messages between them.

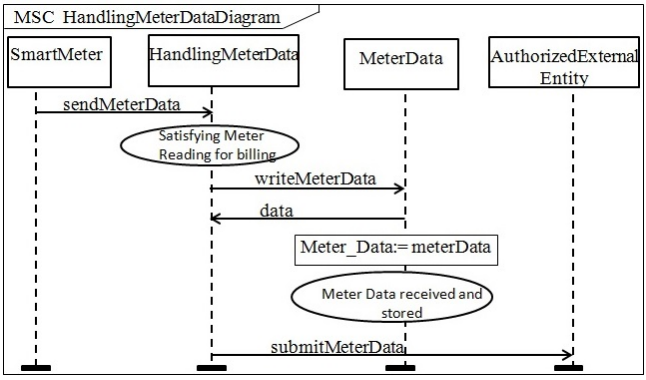


Fig. 7. MSC for the problem diagram related to the goal *Meter Reading for billing*

Step 7 — Refine the Goal Model: Considering the requirements R1 and R4 and security issues, we determine a need for protection of data during transmission. There is no security issue for the requirement R2. The requirement R3 is concerned with storing of data. We therefore need to protect the data during storage. Hence the goal *Meter Reading for billing* is further refined into two soft-goals *Protection of data during transmission* and *Protection of data during storage*. The refined goal model is represented in Fig. 8.

At this point, we reached the Step 7 of the method and continue with the step two by selecting a goal. As our approach is a recursive technique, we select the refined goal *Protection of data during transmission* (**Step 2**) to continue before applying the approach to the next higher level goal (see *Meter registration* in Fig. 5) in the next iteration. As identified in the previous step, the selected goal is related to the requirements R1 and R4. We carry on with the

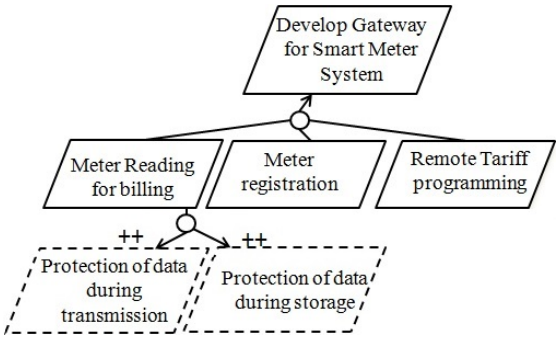


Fig. 8. Goal model and its identified new goals after first iteration of the approach

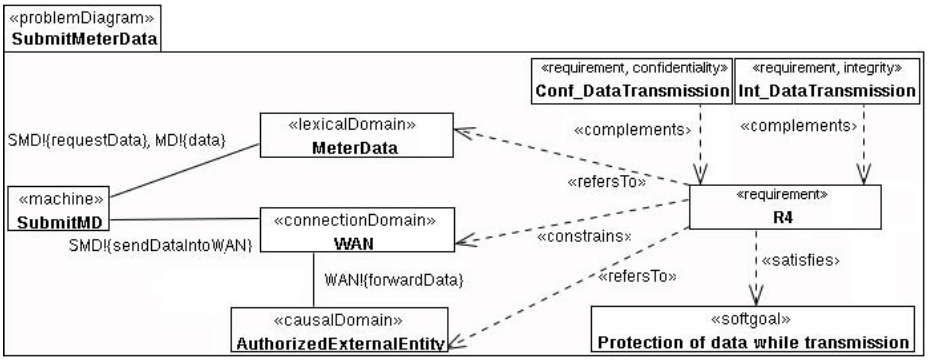


Fig. 9. Linking of the selected refined goal *protection of data while transmission* to the related problem diagram for the requirement R4

next step only for the requirement R4 to identify the relevant domains. From the context of the smart grid depicted in Fig. 4, we identify the new domain *WAN*, when submitting the result to external parties (R4) (**step 3**). We set up the problem diagram for the requirement R4 considering the new identified domain *WAN* (**Step 4**). Figure 9 shows the corresponding problem diagram. It describes that the machine *SubmitMD* submits *MeterData* through the *WAN* to the *AuthorizedExternalEntity*. To achieve the selected soft-goal, we identify two security requirements *Conf_DataTransmission* and *Int_DataTransmission* using the problem diagram from the previous step.

These security requirements state the preservation of confidentiality and integrity of data during transmission. They complement the functional requirement R4 to satisfy the soft-goal *Protection of data while transmission* (see Fig. 9) (**Step 5**). Since the requirements analysis based on the classical problem frames does not support analyzing quality requirements, we extended it by explicitly taking into account quality requirements, which complement functional requirements [21]. We use a UML profile for dependability [10] to annotate problem diagrams with security requirements.

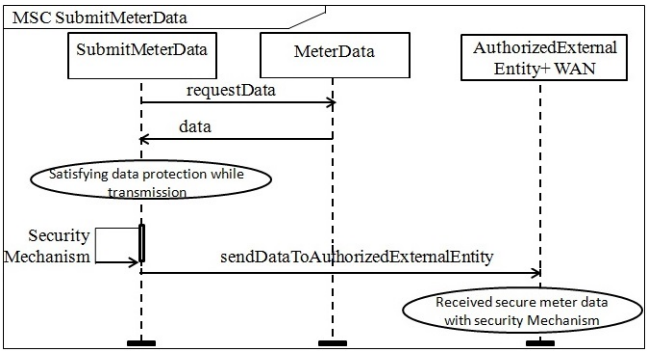


Fig. 10. MSC for the problem diagram related to the soft-goal *protection of data while transmission*

To represent the dynamic behavior of the related context, we set up a MSC shown in Fig. 10 (**Step 6**). To simplify the MSC, we combine the domains *WAN* and *AuthorizedExternalEntity*. The message *sendDataToAuthorizedExternalEntity* represents the phenomena *sendDataIntoWAN* and *forwardData* in the problem diagram 9. Further, a security mechanism is needed to satisfy the security requirements represented in the problem diagram.

At this point, there is no further refinement on the goal model necessary (**Step 7**). Next, we would proceed with Step 2 in the next iteration for the goal *Protection of data during storage*. Ones we treated all the refined goals belonging to the goal *Meter Reading for billing*, we carry on with the next higher level goal *Meter registration*.

Applying our framework to the example of a smart grid system, we have shown how to combine the problem frames approach with a goal modeling technique to use the strengths of each technique and ameliorate the weaknesses. We have analyzed the context of the problem and related it to the goals of the system, explicitly to elicit (quality) requirements systematically. New (soft-)goals have been identified using the knowledge gained from the context analysis. Furthermore, we have used MSCs and related them to the problem diagrams to represent the dynamic behavior of the system-to-be additionally to its static structure represented by the problem frames approach.

5 Related Work

In some work on context modeling, a context diagram is used prior to problem diagrams [6]. The context diagram represents the context by capturing the domains and interconnections of them. A problem diagram contains additionally to the relevant domains, the corresponding requirement. Therefore, the information from the context diagram can also be obtained by using problem diagrams. We use problem diagrams in our approach to represent the relevant context of each problem we want to treat.

There are other problem frame approaches in literature that aim at structuring the solution by fitting the problem to an appropriate problem frame [22]. The authors provide architectural patterns for each problem frame. In this way, they map each problem fitted to a problem frame to one possible solution. The proposed approach is more oriented on the architecture phase. In contrast, our work is concerned with requirements elicitation and their refinement.

A combination of use cases with problem frames is introduced in the work proposed by Del Bianco and Lavazza [23,24]. The authors investigate the possibility of enhancing the problem frames with concepts derived from requirements modeling techniques like use cases based on scenarios and histories. This approach does not cover capturing and modeling the goals of the system. It is similar to the work proposed by Choppy and Reggio [25].

Goal-oriented approaches are integrated within problem frames in [26,27,28]. Bleistein et al. construct a business strategy in a goal model [26,27]. Parts of these goal models are integrated in different problem diagrams instead of requirements. This approach is used to validate system requirements against the constructed business strategy. The proposed approach does not treat the elicitation of requirements and the identification of system goals. The goal models used in these two works are based on i* notation, which makes the requirement part of the problem diagram complex when it comes to defining the requirement.

Liu and Jin use i* goal models and problem frames to enhance the information about the actors in an i* model [28]. The authors map the domain constraints from problem diagrams to i* models. The tasks, which the machine would take over from the actor in a problem diagram are modeled in i* using tasks. The authors do neither consider the elicitation of the requirements nor the definition of goals or mapping them to the context.

6 Conclusions and Future Work

In this paper, we proposed an iterative and recursive approach to bridge the gap between the knowledge of the context and the system purposes. Being aware of the context and relevant domains is essential to avoid extra cost and disastrous situations, in particular when considering security issues of the system. Our approach integrates goal modeling with problem frames and MSCs.

Our main aim is the systematic elicitation of requirements using related goals and relevant domains in the problem context. The problem frame approach enables us to identify the relevant domains in the context of fulfilling a goal. The goals within goal models are mapped to the context of the system. MSCs are used to model the dynamic behavior of the context to satisfy a goal. Furthermore, we provided support for the refinement of goals and soft-goals right up to the elicitation of corresponding requirements, in particular security requirements. Hence, the problem diagrams not only act as a glue for tying a high-level goal to the requirements and the related behavior, but they also support the refinement of goals.

We showed the applicability of our framework using the case study smart grid. We applied our approach on embedded systems with example application of Automatic Cruise Control (ACC) system as well.

The combination of problem frames and goals with support of MSCs provides an innovative step toward analyzing dynamic behavior and static structure of requirements. The overall contribution of our approach can be summarized as follows: (1) eliciting requirements using the related goal and relevant domains in the context; (2) relating requirements - in particular security requirements - to the system goals, in particular soft-goals; (3) identifying new (soft-)goals using problem diagrams; (4) relating dynamic behavior of the system-to-be to the static problem;

We used the UML profile for problem frames (UML4PF), which is conceived as an Eclipse plug in using Eclipse Modeling Framework (EMF) [29]. We enhanced the UML4PF to represent goal notations. The requirements engineer applying our framework has to actively be involved in the process of requirements elicitation and refinement. Therefore, (s)he does not strive for an automatic tool, but for a tool that supports the requirements engineer in applying our approach and facilitate his/her work. In the next step we aim at providing an integrated tool support for the whole approach.

References

1. Nuseibeh, B., Easterbrook, S.: Requirements engineering: a roadmap. In: Proceedings of the 22nd International Conference on Software Engineering (ICSE) on The Future of Software Engineering, pp. 35–46. ACM (2000)
2. Bass, L., Clements, P., Kazman, R.: Software Architecture in Practice. SEI Series in Software Engineering. Addison Wesley (2003)
3. Gross, D., Yu, E.: From Non-Functional Requirements to Design through Patterns. Requirements Engineering 6(1), 18–36 (2001), http://dx.doi.org/10.1007/s007660170013
4. Liu, L., Yu, E.: From Requirements to Architectural Design - Using Goals and Scenarios. In: Proceedings of the 1st International Workshop From Software Requirements to Architectures (SREAW). IEEE Computer Society (2001)
5. Rolland, C., Souveyet, C., Achour, C.: Guiding goal modeling using scenarios. IEEE Transactions on Software Engineering 24(12), 1055–1071 (1998)
6. Jackson, M.: Problem Frames. Analyzing and structuring software development problems. Addison-Wesley (2001)
7. Yang, J., Liu, L.: Modelling requirements patterns with a goal and pf integrated analysis approach. In: Proceedings of 32nd Annual IEEE International Conference on Computer Software and Applications (COMPSAC), pp. 239–246. IEEE Computer Society (2008)
8. Hall, J.G., Rapanotti, L., Jackson, M.: Problem frame semantics for software development. Software & Systems Modeling 4(2), 189–198 (2005)
9. Harel, D., Thiagarajan, P.S.: Message Sequence Charts. In: UML for Real: Design of Embedded Real-Time Systems, pp. 77–105. Kluwer Academic Publishers (2003)
10. Hatebur, D., Heisel, M.: A UML profile for requirements analysis of dependable software. In: Schoitsch, E. (ed.) SAFECOMP 2010. LNCS, vol. 6351, pp. 317–331. Springer, Heidelberg (2010)
11. Côté, I., Hatebur, D., Heisel, M., Schmidt, H.: UML4PF – a tool for problem-oriented requirements analysis. In: Proceedings of the International Conference on Requirements Engineering (RE), pp. 349–350. IEEE Computer Society (2011)

12. Pohl, K.: Requirement Engineering: Fundamentals, Principles, and Techniques. Springer (2010)
13. van Lamsweerde, A.: Goal-oriented requirements engineering: a guided tour. In: Proceedings of the 5th IEEE International Symposium on Requirements Engineering (RE), pp. 249–262. IEEE Computer Society (2001)
14. Yu, E.: Towards modelling and reasoning support for early-phase requirements engineering. In: Proceedings of the 3rd IEEE International Symposium on Requirements Engineering (RE), pp. 226–235 (1997)
15. Nakatani, T., Tsumaki, T., Tamai, T.: Instructional Design of a Requirements Engineering Education Course for Professional Engineers. In: Tsihrintzis, G.A., Jain, L.C. (eds.) Multimedia Services in Intelligent Environments. SIST, vol. 3, pp. 119–151. Springer, Heidelberg (2010), http://dx.doi.org/10.1007/978-3-642-13396-1_6
16. Tsumaki, T., Tamai, T.: A Framework for Matching Requirements Engineering Techniques to Project Characteristics and Situation Changes. In: Proceedings of the 1st International Workshop on Situational Requirements Engineering Processes (SREP). IEEE Computer Society (2005)
17. Jackson, M.: Problem frames and software engineering. Information and Software Technology 47(14), 903–912 (2005)
18. Cox, K., Hall, J.G., Rapanotti, L.: A roadmap of problem frames research. Information and Software Technology 47(14), 891–902 (2005)
19. Kreutzmann, H., Vollmer, S., Tekampe, N., Abromeit, A.: Protection profile for the gateway of a smart metering system. BSI. Tech. Rep. (2011)
20. Requirements of AMI, OPEN meter project. Tech. Rep. (2009)
21. Alebrahim, A., Hatebur, D., Heisel, M.: Towards Systematic Integration of Quality Requirements into Software Architecture. In: Crnkovic, I., Gruhn, V., Book, M. (eds.) ECSA 2011. LNCS, vol. 6903, pp. 17–25. Springer, Heidelberg (2011)
22. Choppy, C., Hatebur, D., Heisel, M.: Architectural Patterns for Problem Frames. In: IEE Proceedings – Software, Special issue on Relating Software Requirements and Architecture (2005)
23. Del Bianco, V., Lavazza, L.: Enhancing problem frames with scenarios and histories in uml-based software development. Expert Systems 25(1), 28–53 (2008), http://dx.doi.org/10.1111/j.1468-0394.2008.00455.x
24. Del Bianco, V., Lavazza, L.: Enhancing problem frames with scenarios and histories: a preliminary study. In: Proceedings of the 3rd International Workshop on Advances and Applications of Problem Frames (IWAAPF), pp. 25–32. ACM (2006)
25. Choppy, C., Reggio, G.: A UML-based approach for problem frame oriented software development. Inf. Softw. Technol. 47(14), 929–954 (2005), http://dx.doi.org/10.1016/j.infsof.2005.08.006
26. Bleistein, S.J., Cox, K., Verner, J.: Problem Frames Approach for e-Business Systems. In: Proceedings of the 1st International Workshop on Advances and Applications of Problem Frames (IWAAPF), pp. 7–15. IEE (2004)
27. Bleistein, S.J., Cox, K., Verner, J.: Validating strategic alignment of organizational IT requirements using goal modeling and problem diagrams. Journal of Systems and Software 79(3), 362–378 (2006)
28. Liu, L., Jin, Z.: Integrating Goals and Problem Frames in Requirements Analysis. In: Proceedings of the 14th IEEE International Conference on Requirements Engineering (RE), pp. 349–350 (2006)
29. Eclipse Modeling Framework Project (EMF), Mai (2013), http://www.eclipse.org/modeling/emf/

Towards a Pervasive Access Control within Video Surveillance Systems

Dana Al Kukhun[1], Dana Codreanu[2], Ana-Maria Manzat[2], and Florence Sedes[2]

[1] Faculty of Computer Informatics, Amman Arab University
Amman - Jordan
danakukhun@gmail.com
[2] Universite de Toulouse – IRIT – UMR 5505,
31062 Toulouse, France
{Firstname.lastname}@irit.fr

Abstract. This paper addresses two emerging challenges that multimedia distributed systems have to deal with: the user's constant mobility and the information's sensitivity. The systems have to adapt, in real time, to the user's context and situation in order to provide him with relevant results without breaking the security and privacy policies. Distributed multimedia systems, such as the one proposed by the LINDO project, do not generally consider both issues. In this paper, we apply an access control layer on top of the LINDO architecture that takes into consideration the user's context and situation and recommends alternative resources to the user when he is facing an important situation. The proposed solution was implemented and tested in a video surveillance use case.

1 Introduction

With the proliferation of multimedia data and applications in a distributed and dynamic environment, many solutions for indexing and accessing multimedia collections are proposed to cope with emerging challenges like: distributed storage and decentralized processing, choice of the indexing algorithms, real time and location-aware information retrieval, optimization of resources consumption (e.g., CPU, RAM, storage, network communications).

In the multimedia contents' management another important issue is raised by the user's mobility and the fact that he/she wants and needs to access the contents from anywhere at any moment: the system's and contents' security. This issue concerns five criteria:

- *Confidentiality*: assurance that the information is shared only among authorized persons or organization;
- *Privacy*: assurance that the identifiable data relating to a person is protected during the information exchange/sharing;
- *Integrity*: assurance that information is authentic and complete;
- *Availability*: assurance that the information is accessible when needed, by those who need it;
- *Traceability*: ability to verify the history, location, or application of an item.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 289–303, 2013.

In this highly dynamic context, the multimedia database systems security becomes a critical issue. Many application domains (e.g., medical, military, video surveillance) may contain sensitive information which should not be or could only partially be accessed by general users. Therefore, it is essential to support security management of multimedia systems and design security models accordingly.

On top of that, we have to consider also that the users are more and more mobile and they need to access the system from anywhere, in real time. There are some cases when the user's context and situation are important, and in order to accomplish a certain task the user must have access to the information. A definition of system security is provided by [1]: "a computing system is secure, if and only if it satisfies the intended purposes without violating relevant informational (or other) rights". Thus, a solution to this privacy and security problem has to be found in order to provide the user with some contents without breaking the system's security.

In this paper, we present a new access control layer on top of the distributed architecture proposed by the LINDO project[1], which considers the user's context and situation within the privacy and security management process. The project's objective was to build a distributed system for multimedia content management, and to ensure effective indexing and storage of data acquired in real time, while considering the resource consumption optimization. The project did not address the issues linked to data privacy and security. The users have full access to all contents after an authentication.

Our objective is to include, within the LINDO framework, the access control in order to attain a pervasive accessibility that enables the user to access multimedia contents at anytime, from anywhere, without breaking the system's security.

In the next section we introduce a state of the art covering distributed access control management, and multimedia access control. The LINDO approach is described in section 3. In section 4, we apply an access control layer on top of the LINDO architecture. In section 5, the adaptive access control solution is illustrated through a video surveillance use case. Finally, conclusions and future work directions are provided in section 6.

2 Related Work

In order to deal with the challenges raised by the big multimedia collections, more and more systems use a distributed architecture for their management. An advantage of this kind of systems is that they benefit from the distributed storage and processing of the contents and the obtained metadata, and thus, a better performance. However, a major problem that these systems encounter is the heterogeneity of indexing algorithms and of the generated metadata and the control of user's access to the system.

In some of the studied systems, complete access is granted after successful authentication of the user. In most application domains a more sophisticated and complex access control is required. In the next section we detail different strategies for distributed access control.

[1] http://lindo-itea.eu/

2.1 Distributed Access Control

In order to guarantee full protection of the confidential information within a decentralized system, accessibility should be controlled through all the communication channels: the application level, the middleware level, the operating system level and finally through the network.

Many models were defined over the years to address the access control issue, without necessarily considering the contents distribution and the user's mobility: DAC [2], MAC [3], RBAC [4], and XACML [5].

In pervasive systems, an important issue that has to be taken into account in the access control management is the user's context and situation in the moment when he/she is accessing the system. More precisely, the system has to react, in real time, to the constant change in the user's context and situation, in order to provide an adaptive access according to the user's needs.

2.2 Context-Aware Access Control

In ubiquitous computing environments, users are mobile and typically accessing resources using mobile devices. As a result, the user's context (e.g., time, location, network connection, device) becomes highly dynamic, and thus, granting him access to the contents without taking his current context into account can compromise the system's security as the user's access privileges do not only depend on "Who the user is", but also on "Where the user is" and "What is the user's state and the state of the user's environment". Thus, access control in ubiquitous applications requires that the user's privileges dynamically change based on his context and role. Thus, permissions assignment for a user has become more complex and dependent on his context.

Many research works have proposed to extend the RBAC model in order to take into account the context's evolution:

- Temporal RBAC [6] considers time as a constraint for the activation and deactivation of a role.
- Spatial RBAC [7] incorporates location information associated with roles in order to permit location-based security policies. Permissions are dynamically assigned to the role dependent on location.
- Dynamic Role Based Access Control [8] dynamically adjusts role and permission assignments based on context information.
- UbiquiTous Role-Based Access Control [9] considers the time and the location of the user as important elements for the activation and disabling of a role. Each role has a state which is changeable during a session.
- [10] provides a context-aware RBAC model that separates context management from the access control model in order to facilitate decision-making in cases where an authorization decision is connected to several contextual constraints.

All these models tie permission assignment to the user's identity and role but also to his contextual attributes, but they are not flexible and responsive enough to deal with any type of situation confronting the user (emergency, un-expected event, etc.).

2.3 Situation-Aware Access Control

The need of situation-awareness in access control was first expressed by [11], who highlighted the importance of providing a security scheme that would relax access rules in order to enable users to meet exceptional circumstances (disasters, medical emergencies or time-critical events).

The works of [12] have also highlighted the importance of providing a flexible security system that would offer more than yes or no answers and that would not rely on predefined solutions in meeting unanticipated access demands.

A flexible solution, called "Break-Glass", was adopted as a standard within health care systems [13]. The solution helps users to confront emergent situations by granting them access to unauthorized needed resources. Despite all the protective steps accompanying it, the "Break-Glass" is an extreme solution that enables users to perform illegitimate intrusions and unjustified access attempts. Therefore, various research works have focused on either controlling the usage of "Break-Glass" by improving its modeling and on facilitating it's integration within the conventional access control models in order to confront the privacy and integrity threats or on proposing other less risky situation-aware access control solutions: [14], [15].

The flexibility level offered by these access control models is directly proportional to the risk of violating the system's security. The more the access control flexibility is performed on a rule-based, predefined or assisted manner, the less the violation risks are introduced. The more the flexibility is provided in an automatic and ad hoc manner, the more the risk level is elevated.

When applying these access control models to the multimedia domain, the things become more complex, especially if a fine-grain access control is needed.

2.4 Multimedia Access Control

Many solutions have been proposed in order to secure the access to multimedia databases and systems. While some authors were interested in the security of the connection to the systems and of the distribution of the contents [16], others were focused on the content-based multimedia access control with fine-grain restrictions at a specific level of the multimedia data [17].

[18] proposes a framework that addresses multi-level multimedia access control by adopting RBAC, XML and Object-Relational Databases. The authors associated roles to users, IP addresses, objects and time periods. All multimedia contents handled by their system have to be segmented. Only the objects which have roles associated to are extracted from the multimedia contents.

[19] studied the confidentiality and privacy issues in the context of a video surveillance system. They defined access rights to different hierarchical objects that can be extracted from the video contents. They focused on the detection of suspicious events.

The management of access control in pervasive environments has evolved over the last years and it takes into account the user's context in the moment when he is interacting with the system. Meanwhile, the distributed multimedia systems do not concentrate their effort on these issues, they consider the management of the contents and their indexing, without the intention of providing solutions for the optimization of the resource consumption.

In order to offer a better response to the user's needs, a system has to leverage between the resource consumption, the returned results and the security. The LINDO project, detailed in the next section, offers a solution for optimal resource consumption, but it does not treat the much attention to the access control.

3 Distributed Multimedia Approach Proposed within LINDO Project

The main goal of the LINDO project (Large scale distributed INDexation of multimedia Objects) is to define a distributed system for multimedia content management, while focusing on the efficient use of the resources in the indexing and query processes. Thus, not only the multimedia contents storage is distributed, but also the indexing process. The originality of this solution is that:

- the content is not moved to indexing servers, but the indexing algorithms are deployed on the remote servers where the content is acquired;
- the indexing process is accomplished in two steps: at acquisition and at query time.

A more detailed presentation of the LINDO architecture can be found in [20].

In order to reduce the resources consumption, the architecture allows the indexing of multimedia contents to be accomplished at acquisition time (i.e., implicit indexing) with some algorithms which extract generic features from the content (e.g., person detection, dominant color detection) and on demand (i.e., explicit indexing) with some algorithms which extract more detailed features (e.g., person recognition, registration plate detection). This avoids executing all the indexing algorithms at once and producing metadata that might never be used, but raises access rights issues concerning the explicit indexing.

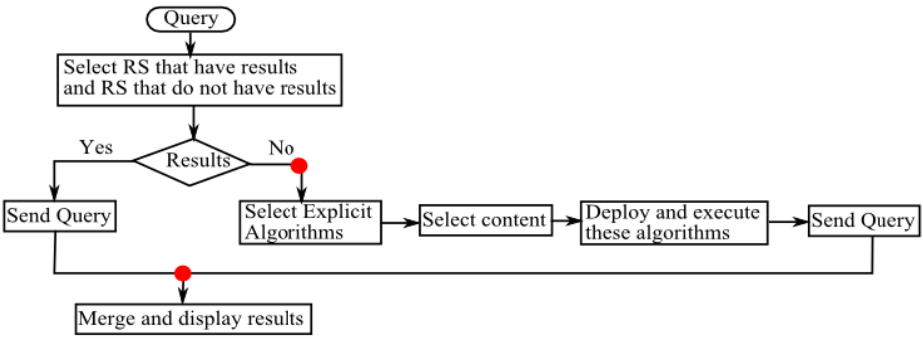


Fig. 1. Query Processing Flow Chart

The query processing (Figure 1) begins with the query specification. First, the query is executed on the metadata collection on the central server, in order to select the remote servers that could provide answers to the query and it is sent for execution to the selected servers. Among the servers that were not selected at the first step, there

could be some servers that contain relevant information that has not been indexed with the right algorithms. For this reason, supplementary algorithms are selected and executed on a sub-collection of multimedia contents. All the results obtained from the remote servers are sent to the central server, where they are combined and displayed to the user.

In this system, the user has full access to all the functionalities and resources after an authentification process. The pervasive access to the system emphasizes the access control and security issues, because of multimedia contents sensitivity and privacy protection laws that impose anonymity constraints. These issues were not treated in the LINDO project.

4 Adding an Access Control Layer to the LINDO Architecture

When applying one of the solutions presented in section 2 to the LINDO system, the lack of results returned to a user's query might not only be due to the lack of results existing within the system but also due to access restrictions imposed by the security layer. This lack of results could be a problem for the user, and prevent him from realizing his task. In order to surpass this limitation, we propose to add an access layer that customizes user's access and is responsible for managing:

The access rights granted to users or services demanding access to multimedia contents that vary according to their role, their context and their situation.

The access rights for executing queries that employ the explicit indexing algorithms: the risk of disclosing personal or confidential information arises with the level of detail sought and provided by the indexing algorithm. For that we have introduced two "checking points" (illustrated in Figure 1, before executing explicit indexing algorithms and before displaying the results) where the user's access rights in his given situation are verified.

In order to achieve these goals, we employ PSQRS (Pervasive Situation-aware Query Rewriting System), an adaptive decision-making system that confronts access denials taking place in real-time situations by rewriting access requests in order to offer alternative-based access solutions, presented in section 4.2. This approach is based on the RBAC model and the XACML standard, and it exploits the user's context (e.g., location, time) and his situation (i.e., the emergency level of the task the user is solving when accessing the system).

The access control relaxation that we propose to carry out respects the access rights defined to protect the multimedia content and applies the adaptive decision-making at two functionalities:

- The explicit indexation execution and the indexing algorithms selection.
- The multimedia contents filtering and presentation.

More precisely, in the two red points illustrated in Figure 1, a matching function is executed, in order to establish based on user's profile and situation if he has the right to run explicit algorithms, respectively to access the content:

$$\Gamma.:(\text{Up, Situation, Policy}) \rightarrow \text{Permit/Deny},$$

where Up is the user's profile, the Situation is a level of emergency (explicitly provided by the user, or inferred from his location, context and other explicit information provided by the user) and Policy is the set of rules that define the security policy of the system.

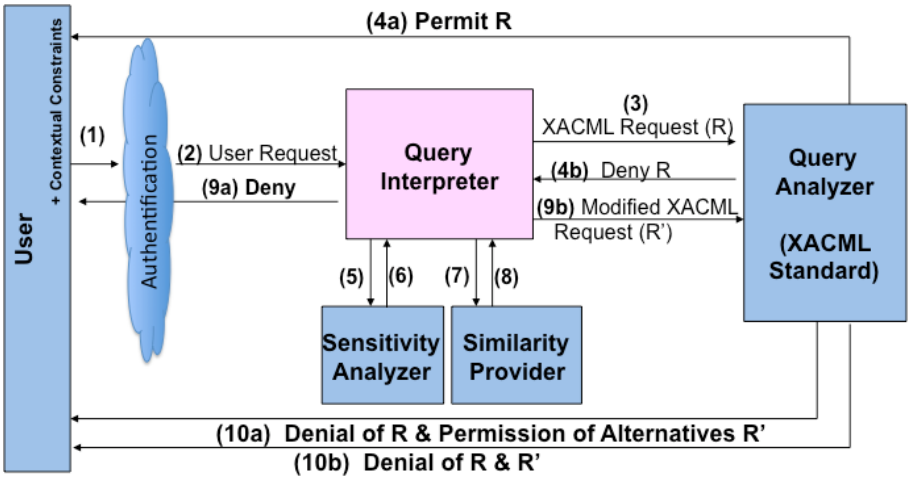


Fig. 2. The PSQRS Architecture

The user profile is defined as:

$$Up = < Uid;Name;Login;Password;RoleId >$$

where Uid is the user's id, Name is the user's name, Login is the user's login, Password is the user's password and RoleId represents the role that is associated to this user for the current access to the system.

An access rule is defined as:

$$Rule = < RuleId;RoleId;Action;Context;Permission >$$

A rule defines a certain Permission (Permit or Deny) for a certain role (i.e, RoleID) and Action (e.g., explicit indexing, object visualization) in a certain Context.

Next, we introduce the detailed functionality of the PSQRS architecture.

4.1 The PSQRS Architecture

As illustrated in Figure 2, the PSQRS architecture contains several components and the sequence of its functionality starts from the user, who enters the system through an authentication portal (step 1) and launches an access request to a certain element (step 2). This request will be interpreted by our Query Interpreter that will translate the request into an XACML request and send it to the Query Analyzer (step3). The request (R) will be analyzed in consideration with the user's profile – automatically extracted at the sign in process -and according to his context. As the analysis finishes,

the Query Analyzer would send the result directly to the user if it's a Permit (step 4a) or back to the Query Interpreter, if it's a deny (step 4b).

In a deny situation the adaptive situation-aware query rewriting mechanism will take place as follows: the Query Interpreter will check the sensitivity of the situation with the help of the Sensitivity Analyzer (steps 5, 6) and according to the situation's importance level, the Query Interpreter will search for similar or alternative resources through the Similarity Provider (steps 7, 8) and employ them to rewrite the XACML request (R') and send it again to the Query Analyzer that will analyze the request and transfer the result back to the user (steps 10a,10b).

4.2 The Fusion of PSQRS and LINDO System

In order to include this access control layer to the LINDO system, some changes have to be done into the system. More precisely, from a functional point of view, each time an access to the system is demanded (the red circles in Figure 1), the PSQRS system is used before applying the explicit indexation and before displaying the results. From an architectural point of view, the system's architecture, detailed in [1], was enhanced with several modules and functionalities.

Thus, in Figure 3, the modules that were added or modified are displayed in red. Almost all the changes are encountered on the central server:

- the Authentification Module was added. This module determines if the user is known by the system and retrieves his profile based on his context. Each user request passes through this module, thus if between two requests the user's context changes his profile could change also.
- the Access Control Module was added. In this module were included the Query Interpreter, the Sensitivity Analyzer and the Similarity provider. All information that is sent to the user passes through this module in order to apply a possible access control adaptation.
- the Terminal Interface was modified in order to capture the user's context and situation.
- a database with RBAC roles, rules and users' profiles was included into the Metadata Engine. It contains also information on the access adaptation.
- the Query Analyzer was integrated into the Request Processor.
- on the Remote Server, only the Access Manager was modified. In fact, a new functionality was added to this module: the execution of an indexing algorithm.

For each identified RBAC role, access rights to the multimedia contents, the explicit indexing and the execution of certain indexing algorithms are specified according to the context the user can have when accessing the system. The user's situation is captured by the system in an implicit (by analyzing the context) or explicit way (in the user interface). In order to offer alternatives to the user according to his situation, the Similarity Provider can select other multimedia contents or execute some algorithms that extract information from the content or modify it in order to respect the user's access rights.

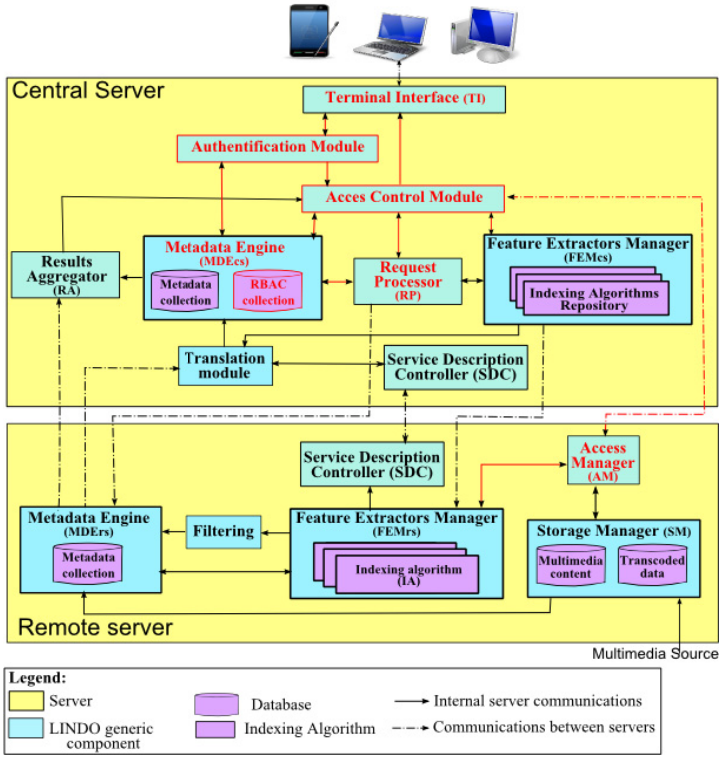


Fig. 3. The modified LINDO Architecture which incorporates the PSQRS approach

In the next section, we present a video surveillance use case, where the implementation of our proposal is used to overcome the lack of answers. As we will illustrate, the system will modify the query processing and will adapt access decisions according to the level of importance of the querying situation.

Table 1. Examples of access rights

Role	**User's Context**	**Content**	**Action**	
			See passenger's faces	**Explicit Indexation**
Security agent	Control room	All	Allow	Allow
	Stations	From metro cameras	Deny	Allow, only object tracking
		From bus cameras	Deny	Deny
Policeman	Control room	All	Allow	Allow
	Stations	From metro cameras	Allow	Allow, only object and person tracking
		From bus cameras	Allow	Deny

5 The Video Surveillance Use Case

This use case concerns a public transportation company that placed surveillance cameras in buses and metros, around the stations and ticket machines.

This system is used by the security agents and police officers. Thus we can identify two roles: security agent and policeman. The access to the system can be done from the control room, or from the stations using a mobile phone. For each role a set of restrictions can be established, based on the existing laws and on the tasks that they have to deal with. Table 1 provides some examples of the access rights given to each role, in a certain context and for a certain multimedia content. For example, a security agent does not have the right to see passenger's faces nor to execute the explicit indexation when he is not in the control room, and he wants to access the multimedia content acquired by the video cameras located in buses.

In this use case we can identify several situations: the security agent investigates on a lost object incident, a lost child research, a bomb attach. Each one of these situations has attached a level of importance (Level=0 is a normal situation, Level=5 is the most important situation).

Let us consider the following scenario: *Taking the bus 2 from "Trocadero" station to "Place d'Italie" station at 14:15, Helen has forgotten her red bag in the bus. As soon as she realized, she went out to report the problem at the information counter in a metro station.*

A typical treatment of such situations goes through the customer service agent who opens a lost object incident with the identification number 1234, takes the descriptions and transmits them to the security agent on site. The security agent will follow different steps in order to find the object. He will check if the object has already been found or returned to the lost and found office by someone. Otherwise, he will execute a query on from his mobile device to check if the object is still in the same location or if somebody took it.

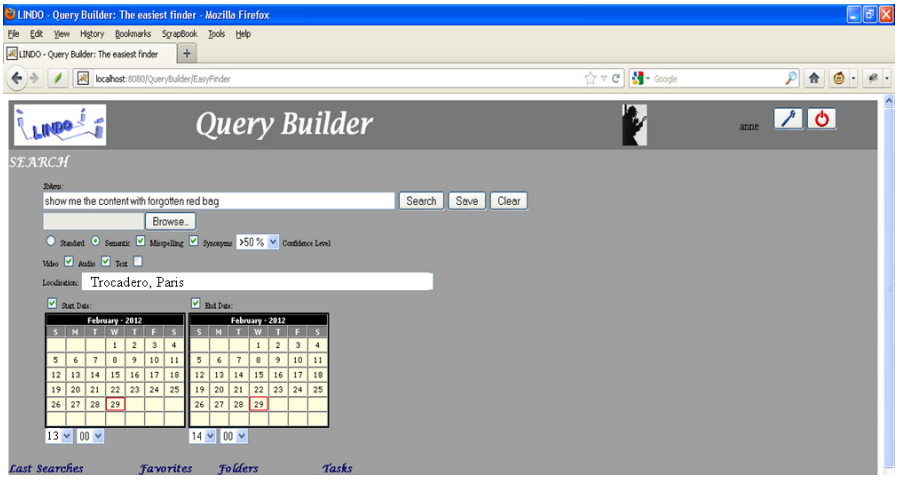


Fig. 4. LINDO user interface

The agent will formulate the following query, in the terminal interface presented in Figure 4 : Find all videos containing a red bag, forgotten in bus nr 2 at Place d'Italie, Paris station, on the 1st of December, between 2:00pm and now (3:00pm), related to the incident number 1234.

In the next sub-sections we present the application of two strategies for ensuring the access control: taking in to account only the user's context, and considering also his situation.

5.1 Applying a Context Access-Control Strategy

The query is processed and translated into XQuery. This query transformation is detailed in [20].

The query processing begins by locating the servers responsible for managing the data captured by the cameras located in the bus nr 2, which passed between 14h00 and 15h00 at Place d'Italie station. Next, a filtering step is performed to restrict the search within the segments captured between 14:00 and 15:00.

The system will then, determine a list of indexing algorithms that would meet the needs and context expressed within the query. Supposing that all the selected algorithms were executed during the implicit indexation process, the query will be executed and thus the video segments that contain red objects are retrieved.

A filtering process is applied to take into account access control rules. Analyzing the access rights assigned to the security agent, we find that he is not authorized to access the videos containing passenger's faces when he is not in the control room. Therefore, considering these access restrictions, the system will eliminate the segments that contain person faces and finally return to the user the list of segments that contain a red object (if available).

5.2 Applying a Situation Access-Control Strategy

The search results returned to the security agent in this case might be insufficient. The red bag might be present in the unauthorized segments containing passenger faces.

An adaptive solution can be employed when the system identifies access challenges related to the user's context or at an important situation. In this scenario, the "lost object" situation is explicitly provided by the agent through the incident number.

The implementation of the adaptive solutions is performed by the PSQRS that adapts decision-making by rewriting the XACML queries.

As shown in Table 2, the richness of the elements that we can embed within an XACML query enables it to describe the contextual attributes characterizing:

- the requested content in the "resource" tag, in red in the figure,
- the user launching the request in the "subject" tag, in bold in the figure
- the situation at which the user has launched the access request in the "environment" tag, in blue in the figure.

The importance level of the situation will determine the level of adaptation.

As the adaptive querying mode is triggered, the query processing mechanism will change to ensure the success of the search by providing a variety of adaptive solutions in correspondence with the situation's sensitivity level.

The adaptation process in this scenario will follow another scheme since the lost object situation is judged to be of higher importance (Level=1). Hence, the Similarity Provider component will be replaced by an Adaptive Solutions Provider. This component will provide some predefined solutions that could bypass the access control challenge or would assist the user in adapting and reformulating his query by pointing out the access challenge and offering him adaptive solutions that would suit his context, the solutions are often saved in a predefined database. Table 3 shows examples of the solutions that the system can offer.

Table 2. XACML request embedding the user's query

```
<Request>
    <Subject>
        <Attribute AttributeId="urn:oasis:names:tc:xacml:2.0:subject:subject-id">
            <AttributeValue>John Smith</AttributeValue> </Attribute>
        <Attribute AttributeId="urn:oasis:names:tc:xacml:2.0:subject:role">
            <AttributeValue>Security Agent</AttributeValue> </Attribute>
        <AttributeAttributeId="urn:oasis:names:tc:xacml:2.0:example:attribute:securityAgent-id">
            <AttributeValue>sa2023</AttributeValue></Attribute>
    </Subject>
    <Resource>
        <ResourceContent>
            <UserQuery>  <QueryInText> Find all videos containing a red bag, forgotten in
bus nr 2 at Place d'Italie station, Paris, on the 1st of December, between 2:00pm and now
(3:00pm)</QueryInText>
                <MediaLocation> bus nr 2 at Place d'Italie station, Paris </MediaLocation>
                <MediaFormat>Video</MediaFormat>
                    <TimeSpan>
                        <From>2012-12-01T14:00:00</From>
                        <To> 2012-12-01T15:00:00</To>
                    </TimeSpan>  </UserQuery>
        </ResourceContent>
    </Resource>
    <Action>
        <Attribute AttributeId="urn:oasis:names:tc:xacml:2.0:action:action-id">
            <AttributeValue>Read</AttributeValue> </Attribute>
    </Action>
    <Environment>
        <Attribute AttributeId="urn:oasis:names:tc:xacml:2.0:environment:environment-id">
            <AttributeValue>Situation</AttributeValue> </Attribute>
        <Attribute AttributeId="urn:oasis:names:tc:xacml:2.0:environment:situation-id">
            <AttributeValue>Forgotten Object</AttributeValue> </Attribute>
        <Attribute AttributeId="urn:oasis:names:tc:xacml:2.0:environment:sitLevel-id">
            <AttributeValue>1</AttributeValue>
         </Attribute>
    </Environment>
</Request>
```

New solutions can also be inserted to the adaptive solutions database through a learning mechanism that detects the solutions that users employ when encountered with access challenges in real time.

The success of the adaptive solutions suggested by the users would eventually be more efficient if they knew the reason behind the access denial. The error messages that often accompany the returned access denial results can serve as indicators to help the users in finding alternative solutions.

Therefore, the adaptive solution for this example will modify the treatment process and will: neglect the filtering step responsible for imposing the access control constraints and replace it with an adaptive step-related to the presentation of resources with unauthorized content.

Table 3. The solutions that our adaptive query processing module can use

Problem	**The adaptive solution**
The privacy law imposing the protection of anonymity of audiovisual contents.	
Passenger faces are not authorized	Display the content after the execution of an algorithm that applies a blur face function.
Voices are not-authorized	Use an algorithm for speech-to-text transcription.
Video volume	
Lack of storage capacity on the user's machine	Use a compression algorithm in order to obtain a smaller file.
Format not supported by the user's machine	Use a conversion algorithm into a compatible format
Download problems due to a low bandwidth	Use a summarization algorithm in order to obtain a concise version of the content.

By applying this process to the scenario described above, the system will return the video segments taken from the Trocadero station between 14:00 and 15:00 and containing a red object. These results will be filtered in order to detect the unauthorized segments (containing passenger faces). This is where the system will apply the adaptation process that would filter the display to conform with the access restrictions imposed by the system.

The adaptation will be performed through a face detection step and the use of an algorithm that applies a "blur function" to protect the privacy of passengers appearing in these segments in order to return to the user a list of relevant results that respect the access rules.

6 Conclusions

In this paper, we have presented an adaptive approach for access control management within multimedia distributed systems, by considering the user's context and situation. Our solution overcomes the access denials that take place in real time access demands by modifying the query processing mechanism and by providing adaptive solutions to bypass the access control constraints. The proposed solution has been validated within the LINDO framework in the context of a video surveillance use case. We applied and validated the same access control approach for other use cases, such as Health care Systems [21].

The adaptive and alternative based situation-aware solution can increase the complexity of processing the request, but if we consider the usefulness of the results

provided in real time and the fact they do not violate the access rights defined by the privacy law, this complexity seems quite acceptable.

In future works, we aim to extend our proposal by taking into account different contextual elements that might also influence the accessibility to multimedia content (e.g., hardware, network bandwidth, etc.) and to apply the adaptive process not only at the presentation level but also at the choice of the explicit indexing algorithms that are protected by RBAC constraints. We plan to exploit users profiles and behavior in order to automatically determine the alternative solutions to use in case when even the PSQRS system returns an access denial. If users do not obtain the desired results, they find other ways to reach their goal. Learning from the experience and the work of others will provide new and interesting adaptive solutions.

Acknowledgments. This work has been supported by the EUREKA project LINDO (ITEA2 – 06011).

References

1. Biskup, J.: Security in Computing Systems: Challenges, Approaches and Solutions, 1st edn. Springer Publishing Company, Incorporated (2008)
2. Harrison, M.A., Ruzzo, W.L., Ullman, J.D.: Protection in operating systems. Commun. ACM 19(8), 461–471 (1976)
3. N.I. of Standards and Technology, Assessment of access control systems, Interagency Report 7316 (2006)
4. Ferraiolo, D., Kuhn, D.: Role-based access controls. In: 15th National Computer Security Conference. NSA/NIST, pp. 554–563 (1992)
5. OASIS, A brief introduction to XACML (March 2003)
6. Bertino, E., Bonatti, P.A., Ferrari, E.: TRBAC: A temporal role-based access control model. ACM Trans. Inf. Syst. Secur. 4(3), 191–233 (2001)
7. Hansen, F., Oleshchuk, V.: SRBAC: A spatial role-based access control model for mobile systems. In: Proceedings of the 7th Nordic Workshop on Secure IT Systems (2003)
8. Zhang, G., Parashar, M.: Dynamic context-aware access control for grid applications. In: Proceedings of the 4th International Workshop on Grid Computing, pp. 101–108. IEEE Computer Society (2003)
9. Chae, S.H., Kim, W., Kim, D.-K.: uT-RBAC: Ubiquitous role-based access control model. IEICE Transactions 89-A(1), 238–239 (2006)
10. Kulkarni, D., Tripathi, A.: Context-aware role-based access control in pervasive computing systems. In: 13th ACM Symposium on Access Control Models and Technologies, SACMAT, pp. 113–122. ACM (2008)
11. Povey, D.: Optimistic security: a new access control paradigm. In: Proceedings of the 1999 Workshop on New Security Paradigms, pp. 40–45. ACM (1999)
12. Rissanen, E., Firozabadi, B.S., Sergot, M.J.: Towards a mechanism for discretionary overriding of access control. In: Christianson, B., Crispo, B., Malcolm, J.A., Roe, M. (eds.) Security Protocols 2004. LNCS, vol. 3957, pp. 312–319. Springer, Heidelberg (2006)
13. Joint NEMA/COCIR/JIRA Security and Privacy Committee (SPC), Break-glass - an approach to granting emergency access to healthcare systems. White paper (2004)

14. Catarci, T., de Leoni, M., Marrella, A., Mecella, M., Salvatore, B., Vetere, G., Dustdar, S., Juszczyk, L., Manzoor, A., Truong, H.-L.: Pervasive software environments for supporting disaster responses. IEEE Internet Computing 12, 26–37 (2008)
15. Kawagoe, K., Kasai, K.: Situation, team and role based access control. Journal of Computer Science 7(5), 629–637 (2011)
16. Sánchez, M., López, G., Cánovas, Ó., Sánchez, J.A., Gómez-Skarmeta, A.F.: An access control system for multimedia content distribution. In: Atzeni, A.S., Lioy, A. (eds.) EuroPKI 2006. LNCS, vol. 4043, pp. 169–183. Springer, Heidelberg (2006)
17. El-Khoury, V.: A multi-level access control scheme for multimedia database. In: Proceedings of the 9th Workshop on Multimedia Metadata, WMM 2009 (2009)
18. Chen, S.-C., Shyu, M.-L., Zhao, N.: Smarxo: towards secured multimedia applications by adopting rbac, xml and object-relational database. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, pp. 432–435. ACM (2004)
19. Thuraisingham, B., Lavee, G., Bertino, E., Fan, J., Khan, L.: Access control, confidentiality and privacy for video surveillance databases. In: Proceedings of the Eleventh ACM Symposium on Access Control Models and Technologies, pp. 1–10. ACM (2006)
20. Brut, M., Codreanu, D., Dumitrescu, S., Manzat, A.-M., Sedes, F.: A distributed architecture for flexible multimedia management and retrieval. In: Hameurlain, A., Liddle, S.W., Schewe, K.-D., Zhou, X. (eds.) DEXA 2011, Part II. LNCS, vol. 6861, pp. 249–263. Springer, Heidelberg (2011)
21. Al Kukhun, D., Sedes, F.: Adaptive solutions for access control within pervasive healthcare systems. In: Helal, S., Mitra, S., Wong, J., Chang, C.K., Mokhtari, M. (eds.) ICOST 2008. LNCS, vol. 5120, pp. 42–53. Springer, Heidelberg (2008)

Analyzing Travel Patterns for Scheduling in a Dynamic Environment

Sonia Khetarpaul[1], S.K. Gupta[1], and L. Venkata Subramaniam[2]

[1] Department of Computer Science and Engg., Indian Institute of Technology, Delhi, India
{sonia,skg}@cse.iitd.ac.in
[2] IBM Research Lab., Delhi, India
lvsubram@in.ibm.com

Abstract. Scheduling a meeting is a difficult task for people who have overbooked calendars and many constraints. This activity becomes further complex when the meeting is to be scheduled between parties who are situated in geographically distant locations of a city and have varying traveling patterns. We extend the work of previous authors in this domain by incorporating some real life constraints (varying travel patterns, flexible meeting point and considering road network distance). We also generalize the problem by considering variable number of users. The previous work does not consider these dimensions. The search space for optimal meeting point is reduced by considering convex hull of the set of users locations. It can be further pruned by considering other factors, e.g., direction of movement of users. Experiments are performed on a real-world dataset and show that our method is effective in stated conditions.

Keywords: Spatio-temporal data mining, geographical Locations, GPS logs, Meeting Points.

1 Introduction

Recent advances in wireless communication and positioning devices like Global Positioning Systems (GPS) have generated significant interest in the field of analyzing and mining patterns present in spatio-temporal data. The pervasiveness of location-acquisition technologies (GPS, GSM networks, etc.) has enabled convenient logging of location and movement histories of individuals. The increasing availability of large amounts of spatio-temporal data pertaining to the movement of users has given rise to a variety of applications and also the opportunity to discover travel patterns. Managing and understanding the collected location data are two important issues for these applications.

The amount of data generated by such GPS devices is large. For example, most GPS devices collect location information for a user every 2 to 5 seconds [8]. This means that for a single user between 17000 to 44000 data points are generated in a single day. Aggregated over tens of users over several days the data size grows exponentially [8,9]. This is extremely rich data and a lot of useful analysis can be performed on this data, potentially giving rise to a variety of application.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 304–318, 2013.

The objectives of this paper is to analyze historical data, determine spatio-temporal relationships among users and predict their behavior efficiently and to determine the optimal meeting point for n users on a road network.

Different applications ranging from location-based services to computer games require optimal meeting point (OMP) query as a basic operation. For example, an educational institute may issue this query to decide the location for a institute bus to pick up the students, so that the students can make the least effort to get to the pickup point. This is also true for numerous other scenarios such as an organization that wants to find a place for its members to hold a conference. This can also be helpful for deciding common meeting points, for social networking site users, having common interests. In strategy games, a computer player may need this query as part of the artificial intelligence program, to decide the appropriate routes.

We introduce two measures to evaluate the processing cost i.e. the minimum-sum-center and the direction of movement. These measures operate over the spatio-temporal domain of each moving objects by applying a network distance to all objects tracked. Each measure induces a spatio-temporal relation that minimizes or maximizes a property over the underlying network graph for the given measure and the given set of moving users. We develop query processing algorithms for computing the value of these measures and to determine spatio-temporal relations and the point on the road network that yields the optimal value of relation's value from the predictive graph of moving objects. Finally, we demonstrate how object movement histories and projected movement trajectories can be used to determine the optimal meeting point.

1.1 Problem Statement

The problem statement can be stated in terms of input and output as follows:

- **Input :** Given GPS logs of n users, proposed meeting time, road network
- **Output:** An Optimal Meeting Point Location(Latitude,Longitude) for given n users

We assume that users have travelling patterns which can be discovered from logs of their travel history. These logs are generated by standard GPS devices. We assume that these devices generates traces at the same rate and in same format(same granularity) or GPS traces can be transformed to a fixed format. We define the optimal meeting point for group of people as:

Let $\{u_1, u_2, ..., u_n\}$ be a set of n users. The users have a definite travel pattern(may be periodical) which is hidden in their logs. Let $u_i(t)$ denotes the location of user i at time instant t. It is noted that if all of these n users wish to meet at some fixed time, and the meeting point is desired as OMP, then the sum of distances would be the total cost of meeting(for all the users). It is proposed to minimize this cost.

Optimal meeting point is defined to be $arg\ min_{OMP \in N}[\Sigma d_N(u_i(t), OMP)]$, where $d_N(x, y)$ is the shortest distance between two point x and y on the road network N.

Informally, we define the **optimal meeting point** as a point on the road network where the sum of distances travelled by all the users is minimum. In Figure 1 total distances travelled by all users is minimum at point OMP, and is 23 K.M.

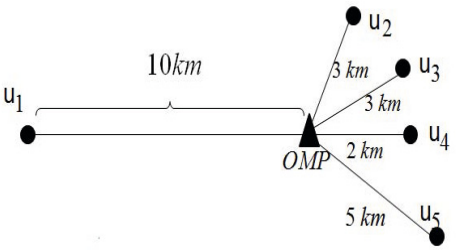


Fig. 1. Optimal Meeting Place

We also consider predicted directions of motion of users at time t and at time $t + \delta t$ to optimize the total distance covered by each user. It is considered by examining the consecutive hulls of of location points.

The rest of this paper is organized as follows. In the following subsection we describe related work and also put the work of this paper in the context of related work. Later in section 2, we introduce algorithms, describing the underlying idea to determine the optimal meeting point. In Section 3, we report the results of our experiments on real-world dataset to show the feasibility of our algorithms for the OMP queries. Finally, we conclude our paper in Section 4.

1.2 Related Work

There has been a lot of prior work on using spatio-temporal data to track movement history. There has also been work around integrating multiple users' information to learn patterns and understand a geographical region. GeoLife is a location-based social-networking service on Microsoft Virtual Earth. GeoLife enables users to share travel experiences using GPS trajectories [3, 6, 8, 9, 14]. It finds the top most interesting locations, classical travel sequences in a given geospatial region. In geolife to find out interesting locations in a given geospatial region HITS model (Hypertext induced topic model search) is introduced. The paper [7] is based on Hybrid Prediction Model, which estimates an object's future locations based on its pattern information as well as existing motion functions using the object's recent movements. Other systems like and CityVoyager [10] are designed to recommend shops and restaurants by analyzing multiple users' real-world location history. Mobile tourist guide systems [6, 10–12] typically recommend locations and sometimes provide navigation information based on a user's real-time location. In contrast, our approach is based on assumption that users are moving not stationary, and follows a regular routine during weekdays. Our approach to determine users location points is based on simple statistics that applied on users historical GPS traces.

OMP problem is well studied in different forms in Euclidean space. In Euclidean space optimal meeting point is basically the Geometric median of location point set. When the Euclidean distance is adopted as the metric of distance, the OMP query is called the Weber problem [17], and the OMP is called the geometric median of the query point. Like various nearest neighbor queries [5, 13, 15, 16], the OMP query is also fundamental in spatial databases.

In our previous work [1] we have determined common meeting point from spatio-temporal graph analysis for two users in the Euclidean space. In this paper we determine the optimal meeting point from the trajectory analysis for n users on road network. This brings in several factors in the analysis, making the system flexible and more realistic to use, while adding complexity to analysis.

On the other hand, the OMP query is not well explored in terms of road networks, where the network distance is adopted as the distance metric. However, compared with the Weber problem, this is a more realistic scenario for location-based services. Recently, [4] proposed a solution to this problem by checking all the split points on the road network. It is proved in [4] that an OMP must exist among the split points, which leads to an algorithm that checks the split point of each query point in Q on each edge in the road network G = (V, E), and picks the split point with the smallest sum of network distances as the OMP. As a result, the search space is $|Q|.|E|$, which is huge. Although [4] includes a pruning technique to skip some split points that are guaranteed not to be an OMP, the search space after pruning is still very large. Therefore, a novel road network partitioning scheme is proposed in [4] to further prune the search space, based on the property that the OMP is strictly confined within the partition where all the objects in the query set Q are located. After that [2] proves that an OMP must exist either on vertex or on query points, there is no need to check all the split points. So search space is reduced to $|Q| + |V|$. To further reduce the search space two phase convex-hull-based search space pruning techniques are proposed in [2]. In contrast our approach considers that user are moving not stationary, so we are predicting user locations, the directions in which they are moving and pruning the search space based on their locations before and after the meeting. Our approach considers both spatial and temporal aspect of data. Our approach is defined in following section.

2 Determining Optimal Meeting Points for Multiple Users

GPS and other positioning devices generate location information every few seconds (often at the interval of two to five seconds). An individual carrying such a device potentially generates thousands of GPS points everyday. It is important to be able to aggregate all the data from multiple users and predict their location points at the given time. In this paper we will apply statistical operation to generate spatio-temporal location point prediction for each user at given time. After predicting the location points for the user optimal meeting point is determined.

We will now define the notation and also describe the problem that we are solving. We assume that there are n users whose time stamped GPS logs are available to us.

GPS Users: We have GPS logs of n users U = $\{u_1$,u_2,.....,$u_n\}$.

GPS Point: A GPS point g_i is a four field tuple, $\langle x, y, d, t\rangle$ where x, y are geographic coordinates (Latitude, Longitude respectively) and d, t is the timestamp (date, time respectively) represents a user's location at any point of time.

The pair (Longitude, Latitude) represents the position of the user at a particular point of time. For the purposes of standardization, we assume that each value in the pair is

given in six decimal places. A value in decimal degrees to 6 decimal places is accurate to 0.111 meter at the equator [20].

GPS Trajectory: A GPS trajectory Tr_i is a sequence of GPS traces ordered by timestamp $Tr_i = (g_{i_1}, g_{i_2}....., g_{i_n})$ of the user i.

We divide our region of interest into a grid C of $m * n$ cells where m is number of unique δLat and n is number of unique δLng.

Cell: A cell c_i is a rectangular element of grid C dividing the region of interest. They are sequenced major rowwise and minor columnwise.

Road Network Distance: Road network distance is the shortest length of path between two cells on road network. This distance is obtained using function $d_N(c_i, c_j)$, it returns the length of path between two cells of grid C, c_i and c_j on road network. This distance function $d_N(c_i, c_j)$ can be realized through google maps.

Location Point: A location point l_i represents the location of the i^{th} user on grid C at a point of time.

We determine the location points of all individual users at given point of time. Location point of a user is a predicted geographical location where the user is, at given meeting time. This prediction is done by statistical analysis of their past GPS logs.

2.1 Location Point Determination

Most of us generally follow a specific travel pattern during working days. To determine the location point for each individual user, at a given point of time, we analyze their past GPS logs. By applying statistical operations on their past GPS trajectories, we are able to predict their locations at a given point of time. For location points analysis w.r.t. to time and space, the whole geographical space is divided into grid, where each cell $c(l * w)$ represents a small geographical region and is assigned a number. Twenty four hours in a day are divided into small time periods of length δt. The log records are mapped on to this grid. For each user, his/her location cell number, after every δt interval of time is identified from his/her log records. User locations of many days at different time intervals are summarized to generate his/her spatio-temporal graph.

We determine the maximum and minimum value of latitude and longitude($Ar = (maxlat - minlat) * (maxlng - minlng)$ gives total area of city), which define our domain of interest. $\delta lat * \delta lng$ form the area of a single cell within the grid. They are sequenced major rowwise and minor columnwise, from 1 to K, where K maximum number of cells. Location of each user is predicted in terms of a cell number. Mapping of user's GPS location into cell number and cell number into GPS location is done using following conversions.

Mapping given location (Lat_i, Lng_i) **to cell number:**
$Cellno = ((Lng_i - minLng) / \delta Lng) *$ No. of unique $\delta clng + ((Lat_i - minLat) / \delta Lat)$

Mapping given cell number to location (Lat_i, Lng_i)**:**
$Cellno = lng_{ind} *$ No. of unique $\delta lng + lat_{ind}$

Where lng_{ind} is quotient and lat_{ind} is remainder when $cellno/$ (No of unique δlng)
$Lat_i = minLat + lat_{ind} * \delta Lat$
$Lng_i = minLng + lng_{ind} * \delta Lng$

Using these mapping we are able to plot user historical traces onto grid after every δt time interval. After plotting the user traces on the grid we apply statistical mode operation (defined below) to determine the user location at given time t.

Temporal Mode of User Location Points
The temporal mode of a set of data points is the value in the set that occurs most often during a specified time interval. The mode of i^{th} user's historical data points $Mode(U_i)$ is the value that occur more frequent within a specified the time. For applying the mode operation, users longitude and latitude values are mapped onto the grid so that it points to the defined geographical region. The mode operation is applied for each user.

Let U_i be the set of GPS points of user i at time δt and l_i be the most frequently visited location point of user for given time δt.

$l_1 = Mode(U_1)$
$l_2 = Mode(U_2)$
$l_n = Mode(U_n)$
$L = \{l_1, l_2, l_3, ..., l_n\}$

For a set of users, we can determine the cell numbers(location points) in which they are expected to be at any point of time. Let the set of location points of n users is denoted with $L = \{l_1, l_2, ..., l_n\}$ at time t.

By applying these statistics we determine the cell where an individual user is mostly present at a given point of time. We discard the cells which are visited very few number of times in long period of GPS traces.

A baseline algorithm to solve the meeting point problem is defined below.

2.2 Baseline Algorithm

For a given set of users, let L be union of set of location points. The baseline algorithm considers all the cells $|C|$ within the grid as the probable candidates for an optimal meeting point. The baseline algorithm evaluates the sum of distances from the location points of each user to each cell and the OMP is the cell with the minimum value of the sum.

The approach is presented in Algorithm 1 where function d_N computes shortest distance between a location point l_i and cell c_j on road network using any standard procedure (in our case Google Maps API).

The baseline algorithm has very high computational complexity $O(nK)$, where n number of users and K is the number of cells, as it considers the entire grid as the search space. It is required to prune the space to overcome this problem. We propose the use of a two level convex hull pruning to reduce the search space, and there by improving the efficiency.

Algorithm 1. BaseLine Algorithm(L,C)

Data: Location points $L = \{l_1, l_2, \ldots, l_n\}$ of n users at given time T_i on the Grid containing $|C|$ cells
Result: Optimal Meeting Place- A cell on the Grid $OMP(Lat, Lng)$
begin
  $OMP \leftarrow NULL$
  $mincost \leftarrow +\infty$
  foreach $c_i \in C$ **do**
    $sum \leftarrow 0$
    foreach $l_j \in L$ **do**
      $sum \leftarrow sum + d_N(c_i, l_j)$
    $cost \leftarrow sum$
    if $cost < minCost$ **then**
      $mincost \leftarrow cost$
      $OMP \leftarrow c_i$
  $Return(OMP)$

So baseline approach is to consider all the cells on the grid as a search space to determine the optimal meeting point. In our approach we are using two level convex hull pruning to reduce the search space and to improve the efficiency of search.

2.3 Convex Hull Based Pruning

The convex hull $H(L)$ of a set L is the intersection of all convex sets of which L is a subset. It is also the union of all straight lines joining all pairs of points in L [19].

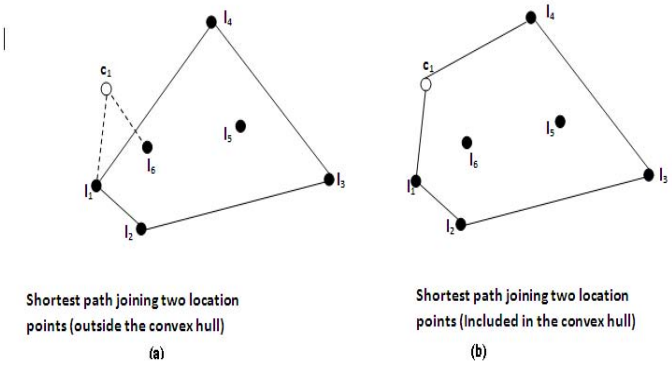


Fig. 2. Counter example

It can be observed that given a set of location points L, a minimum distance point from all location points of set L, i.e., $argmin_{x'}[\Sigma d_E(l_i, x')]$ always lies inside the convex hull H(L), where function $d_E(x, y)$ returns Euclidean distance between points x and y. It can be deduced from the property of a convex object that its centroid lies within the object [19]. But, as shown in figure 2(a), it may not be always true for road network.

To ensure this property for road network, we calculate the shortest route between every two location points using function $shortRoute(x, y)$, all the points that lies among the routes are merged with location points set as described in figure 2(b). After that we take the convex hull of this set.

According to the baseline algorithm, OMP must exist among one of cells in the Grid. It is not necessary to check all the cells in the grid. The search space can further be pruned. We check only those cells that are in the smallest partitioned grid enclosing all cells of the user's location points. We define a convex hull based pruning technique in Algorithm 2, where $convexHull(L)$ computes the convex hull of the point set L using Andrew's Monotone Chain algorithm [18] and takes $O(|P|log|P|)$ time where P is the number of points. All the cells those lie within the hull, we collect them into set P. Now to determine the OMP we check only those that are belong to set P.

Algorithm 2. ConvexHullPruning(LocationPoints,Cells)

Data: Location points $L = \{l_1, l_2, \ldots, l_n\}$ of n users at given time T_i on the Grid consists of cells C
Result: Set of Grid Cells P lying inside the Convex Hull
begin
  $OCells \leftarrow 0$
  foreach $l_i \in L$ **do**
    foreach $l_k \in L$ **do**
      $OCells \leftarrow shortRoute(l_i, l_k)$
  $L_n \leftarrow L \cup OCells$
  $P \leftarrow 0$
  $H \leftarrow ConvexHull(L_n)$
  foreach $c_i \in C$ **do**
    if $c_i \in H$ **then**
      $P \leftarrow P \cup c_i$
    else
      Discard c_i
  $Return(P)$

The search space is significantly reduced using convex hull based pruning. It is further possible to trace the direction of movement from the user data. We can further reduce the search space by considering the direction of movement of the Users'.

2.4 Direction of Movement Based Pruning

Let $L(t) = \{l_1, l_2, \ldots, l_n\}$ be the set of location points of users at time t and let P be the set of cells of the grid lying within the convex hull of $L(t)$. Similarly, let $L'(t+\delta t) = \{l'_1, l'_2, \ldots, l'_n\}$ be the set of location points at time $t + \delta t$ and P' be the set of cells of the grid lying within the convex hull of $L'(t + \delta t)$. By analyzing the two consecutive convex hulls, it is found that two cases are possible.

Case 1: The two convex hulls are intersect or overlap.

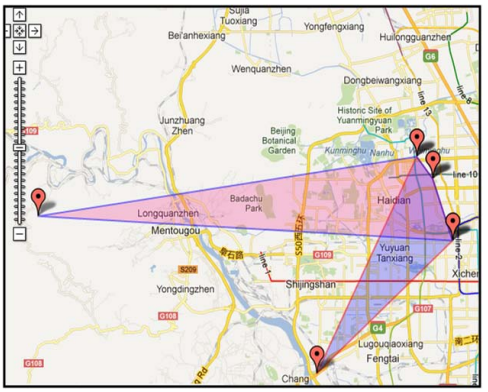


Fig. 3. Intersecting Hulls

In this case, the meeting point among the cells is assumed to lie inside the intersection/overlapped region. Figure 3 depicts the intersecting and figure 4 depicts the overlapped convex hulls of four users. It may be noted that this case also covers the case if a convex hull is completely contained in the other convex hull shown in figure 5. In this case, we assume that the meeting point would lie in the smaller convex hull. We take meeting point inside the intersection because it reduces the sum of total distance travelled by users before and after the meeting.

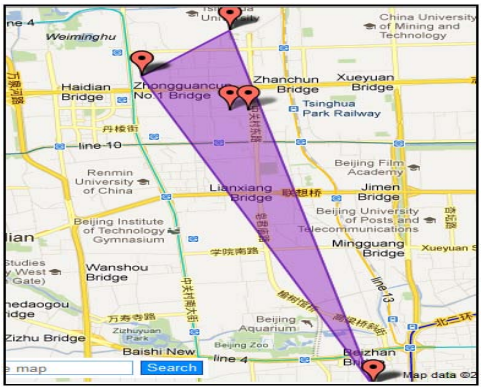


Fig. 4. Overlapped Hulls

Case 2: The two convex hulls are disjoint of each other (i.e. they have zero intersection). In this case, we assume that the meeting point would lie in the first convex hull. We take meeting point inside the intersection because it reduces the sum of total distance travelled by users to reach the meeting point.

These two levels of convex hull pruning prune the search space to P'', a considerably reduced set. A complete algorithm 4 is developed to determine optimal meeting point in such a scenario. Function Map(OMP,OMPLoc) converts cell designated as OMP into (Latitude,Longitude) pair represented by OMPLoc.

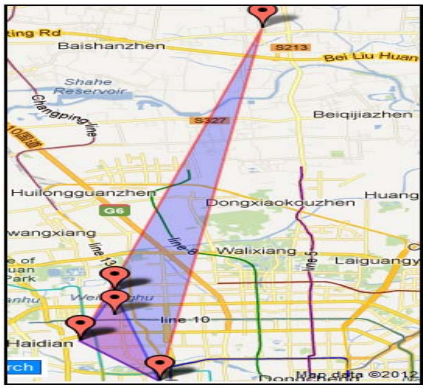


Fig. 5. Inner Hull

Algorithm 3. DirectionPruning(LocationPoints,Cells)

Data: Set of Grid Cells P at given time T_i and P' at time $T_i + \delta t$
Result: Set of Grid Cells P'' selected after pruning
begin
  $P'' \leftarrow P \cap P'$
  if $P''isNull$ **then**
    $P'' \leftarrow P$
  $Return(P'')$

3 Experiments

In this Section, we first present details about the GPS dataset used [6, 8, 9]. Next we present the results of applying statistical operations on the temporal aspect of GPS data for predicting user location points at a given time. Then, as stated, two level of pruning are used to determine OMP.

We compare the results of our approach with two other approaches in terms of total distance travelled by users before and after the meeting and number of cell searched. First approach is when we do not consider the direction of movement of users and second approach is when we consider all location points before and after the meeting i.e. L and L' simultaneously.

3.1 GPS Trajectory Dataset

The GPS trajectory dataset [6, 8, 9] is a repository of real life data collected by Microsoft Research. The data was collected by 165 users in a period of over two years (from April 2007 to August 2009). A GPS trajectory of this dataset is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude, height, speed and heading direction, etc. These trajectories were recorded by different GPS loggers or GPS-phones, and have a variety of sampling rates.

Algorithm 4. OMP Algorithm(L,L',C)

Data: Location points $L = \{l_1, l_2, \ldots, l_n\}$ and $L' = \{l'_1, l'_2, \ldots, l'_n\}$ of n users at given time T_i and $T_i + \delta T_i$ on the Grid G containing $|C|$ cells
Result: Optimal Meeting Place- A cell on the Grid $OMPLoc(Lat, Lng)$
begin
  $P \leftarrow ConvexHullPruning(L, C)$
  $P' \leftarrow ConvexHullPruning(L', C)$
  $P'' \leftarrow DirectionPruning(P, P')$
  $mincost \leftarrow +\infty$
  foreach $c_i \in P''$ **do**
    $sum \leftarrow 0$
    foreach $l_j \in P''$ **do**
      $sum \leftarrow sum + d_N(c_i, l_j)$
    $cost \leftarrow sum$
    if $cost < minCost$ **then**
      $mincost \leftarrow cost$
      $OMP \leftarrow c_i$
  $Map(OMP, OMPLoc)$
  $Return(OMPLoc)$

95 percent of the trajectories are logged in a dense representation, e.g., every $2 \sim 5$ seconds or every $5 \sim 10$ meters per point, while a few of them do not have such a high density being constrained by the devices. This dataset recorded a broad range of users' outdoor movements at different time intervals over a day, including not only life routines like going home and going to work but also some entertainment and sports activities, such as shopping, sightseeing, dining, hiking, and cycling etc. In the data collection program, a portion of users carried a GPS logger for more than two years, while some of them may have carried a logger for few weeks.

We worked on 126 users from the above dataset and worked on their GPS traces. This subset consists of a total of 68612 days data with 5,832,020 GPS points. The total area covered by the GPS logs exceeded 3,880,951 Sq. kilometers. The majority of the data was created in Beijing, China. A large part also came from Hangzhou. We have partitioned the total area covered by users into grid. Following subsections explain the grid formation for further processing.

3.2 Determining Grid

The major portion of this dataset belongs to Beijing, China. A grid is defined over this area, which is approximately 38972.068 Sq. kilometers, with minLat=38.0, minLng =115.0 and maxLat=40.0, maxLng=117.0. This is further divided into small cells of (222 meter * 219 meter) area with δLat=0.002 and δLng=0.025. Thus we have a total of 1,80,00,00 cells. These cells are sequenced major rowwise, minor columnwise and identified by unique numbers from 1 to 1800K. Users GPS points are mapped into these cells and their location points are predicted.

After partitioning the city into the grid, the aim is to compute the user locations at a given point of time and map them onto grid.

3.3 Predicting User Location Points

Recent one month historical data of a user is analyzed to predict his/her location at a given time. For example if we have to predict the user location for 18 April at 10 am. then we extract his last month data points for the time interval 9:45 to 10:15 am. Data points are mapped on to grid and temporal mode operation is applied to determine the cell with maximum frequency of data points. This cell is marked as location point for the user at 10 am.

This experiment is performed for hundreds times and prediction is made for days for which data is already available. It is observed that 73% of times predicted cell is same as actual location of user. For all prediction made a **mean square error of 0.238** was determined. Figure 6 shows a mean square error detected for 104 prediction made.

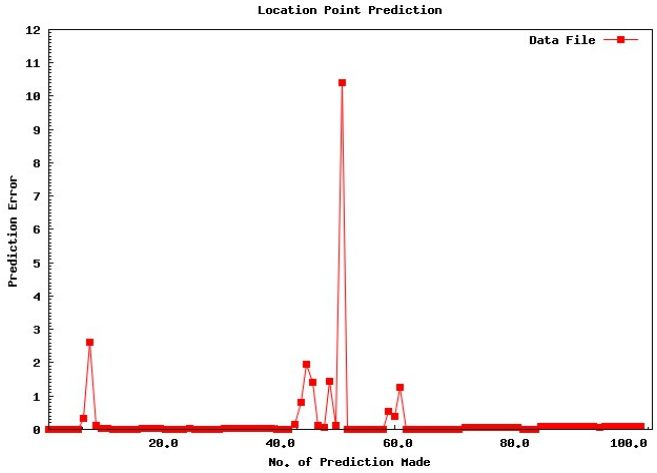


Fig. 6. Prediction Error

After computing users locations our aim is to determine a meeting point for them on road network.

3.4 Determining Optimal Meeting Place

To check the efficiency and accuracy of our approach, we conducted experiments to determine optimal meeting point in three different ways, described below.

OMP by Considering set *L*(No Directions). In this case, we determine the meeting point, for a given meeting time, by considering only set L. For example, if the meeting time is at 10 am, then search space is pruned by considering users' location at 10 am only. A convex hull of current location points L is calculated and the optimal meeting point is determined within this convex hull.

OMP by Considering Set L and L' Simultaneously. In this case, we determine the meeting point, for a given meeting time, by considering both set L and L' at the same time. For example, if meeting time is at 10 am and meeting duration is one hour, then we prune the search space by using both sets of user locations at 10 am and at 11 am simultaneously. Convex hull of all location points is calculated and the optimal meeting point is determined within this convex hull.

OMP by Considering Direction of Movement of Users with Time. In this case, we consider the direction of movement of users to determine the meeting point. For example, if meeting time is 10 am and meeting duration is one hour, then we prune the search space by considering users' location points at 10 am and at 11 am. We first determine the convex hull of locations points at 10 am and then we determine the convex hull of user locations at 11 am. An optimal meeting point is determined within the intersecting/overlapped area of two convex hull.

A table 1 summarizes the average distance travelled by all users and average number of cells to be searched in each of these three cases is given below.

Table 1. Distance Travel and Cells to be searched in 3 different Cases

Without Directions		All Location Points		With Directions	
$Avg.Distance_1$	$Avg.Cells_1$	$Avg.Distance_2$	$Avg.Cells_2$	$Avg.Distance_3$	$Avg.Cells_3$
35.032	454.925	31.00505	407.375	31.94225	74.125

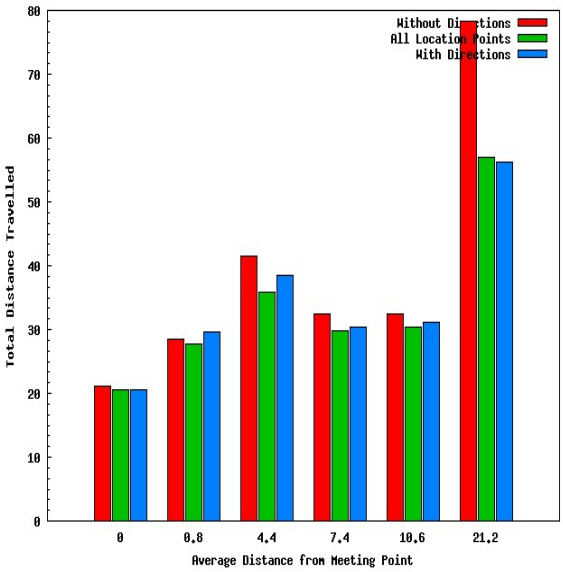


Fig. 7. Distance Travelled by users

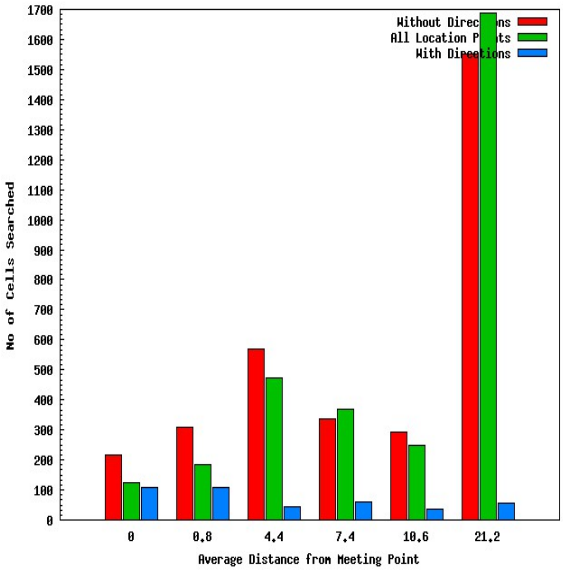


Fig. 8. Search Space reduction

Experiment is performed 50 times for different number of users and Optimal meeting point is determined in each of these three cases. Figure 7 shows the graph of total distance travelled by users in each of these three cases. The X-axis shows the average distance travelled by users. Graph shows that users have to travel much more distance in case 1, when we do not consider direction of movement of users. In case 2 and case 3 these distances are nearly same.

Figure 8 shows the graph of number of cells to be searched to determine the optimal meeting point in each of these three cases. Graph shows that more number of cells are required to be search in case 1 and case 2 as compare to case 3. It shows that the search space is significantly reduced if we apply the two-level convex hull pruning.

4 Conclusions

Scheduling a meeting is a difficult task for people who have overbooked calendars and many constraints. The complexity increases when the meeting is to be scheduled between parties who are situated in geographically distant locations of a city and have varying travel patterns.

In this paper, we investigated the problem of *identifying a common meeting point for a group of users who have temporal and spatial locality constraints that vary over time*. We solved the above problem for a number of users on road network by using the GPS traces of the users.

We begin by mining historical GPS traces of individual user and mapped their locations onto predefined grid and predicted their location points for the given meeting time. We then applied two levels of convex hull based pruning to reduce the search

space and determine the optimal meeting point. We have used predicted future direction of movement of the users to reduce total distance travelled by users before and after the meeting.

The method was evaluated on a large real-world GPS trace dataset and showed the effectiveness of our proposed method in identifying a common meeting point for an arbitrary number of users on the road network.

References

1. Khetarpaul, S., Gupta, S.K., Subramaniam, L.V., Nambiar, U.: Mining GPS traces to recommend common meeting points. In: Proceedings of IDEAS 2012, pp. 181–186 (2012)
2. Yan, D., Zhao, Z., Ng, W.: Efficient Algorithms for Finding Optimal Meeting Point on Road Networks. In: Proceedings of the VLDB Endowment, pp. 968–979 (2011)
3. Khetarpaul, S., Chauhan, R., Gupta, S.K., Subramaniam, L.V., Nambiar, U.: Mining GPS Data to Determine Interesting Locations. In: Proceedings of IIWeb 2011, WWW 2011 (2011)
4. Xu, Z., Jacobsen, H.-A.: Processing Proximity Relations in Road Networks. In: Proceedings of SIGMOD, pp. 243–254 (2010)
5. Chen, Z., Shen, H.T., Zhou, X., Yu, J.X.: Monitoring Path Nearest Neighbor in Road Networks. In: Proceedings of SIGMOD, pp. 591–602 (2009)
6. Zheng, Y., Zhang, L., Xie, X., Ma, W.: Mining correlation between location using human location. In: Proceedings of ACM GIS 2009, pp. 625–636 (November 2009)
7. Jeung, H., Liu, Q., Shen, H.T., Zhou, X.: A Hybrid Prediction Model for Moving Objects. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering (ICDE 2008), pp. 70–79. IEEE Computer Society, Washington, DC (2008)
8. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Geolife: Managing and understanding your past life over maps. In: Proceedings of MDM (April 2008)
9. Zheng, Y., Zhang, L., Xie, X., Ma, W.-Y.: Understanding mobility based on gps data. In: Proceedings of Ubicomp, pp. 312–321 (September 2008)
10. Simon, R., Frohlich, P.: A mobile application framework for the geospatial web. In: Proceedings of WWW, pp. 381–390 (May 2007)
11. Beeharee, A., Steed, A.: Exploiting real world knowledge in ubiquitous applications 11(6), 429–437 (2007)
12. Park, M., Hong, J., Cho, S.: Location-based recommendation system using bayesian user's preference model in mobile device. In: Proceeding of UIC, pp. 1130–1139 (July 2007)
13. Mouratidis, K., Yiu, M.L., Papadias, D., Mamoulis, N.: Continuous Nearest Neighbor Monitoring in Road Networks. In: Proceedings of VLDB, pp. 43–54 (2006)
14. Horozov, T., Narasimhan, N., Vasudevan, V.: Using location for personalized poi recommendations in mobile environments. In: Proceedings of SAINT, pp. 124–129 (January 2006)
15. Cho, H., Chung, C.: An Efficient and Scalable Approach to CNN Queries in a Road Network. In: Proceedings of VLDB, pp. 865–876 (2005)
16. Yiu, M.L., Mamoulis, N., Papadias, D.: Aggregate Nearest Neighbor Queries in Road Networks. IEEE Trans. on Knowl. and Data Eng. 17(6), 820–833 (2005)
17. Cooper, L.: An Extension of the Generalized Weber Problem. Journal of Regional Science 8(2), 181–197 (1968)
18. Preparata, F.P., Shamos, M.I.: Computational Geometry:An Introduction. Springer (1985)
19. Convex Hull and properties (March 28, 2013), `http://en.wikipedia.org/wiki/Convex_set`, `http://en.wikipedia.org/wiki/Centroid`
20. Accuracy versus decimal places at the equator (March 28, 2013), `http://en.wikipedia.org/wiki/Decimal_degrees`

Human–Computer Interaction and Knowledge Discovery (HCI-KDD): What Is the Benefit of Bringing Those Two Fields to Work Together?

Andreas Holzinger[1,2]

[1] Medical University Graz, Institute for Medical Informatics, Statistics and Documentation
Research Unit Human–Computer Interaction
Auenbruggerplatz 2/V, A-8036 Graz, Austria
a.holzinger@hci4all.at
[2] Graz University of Technology, Institute for Information Systems and Computer Media
Inffeldgasse 16c, A-8010 Graz, Austria
a.holzinger@tugraz.at

Abstract. A major challenge in our networked world is the increasing amount of data, which require efficient and user-friendly solutions. A timely example is the biomedical domain: the trend towards personalized medicine has resulted in a sheer mass of the generated (-omics) data. In the life sciences domain, most data models are characterized by complexity, which makes manual analysis very time-consuming and frequently practically impossible. Computational methods may help; however, we must acknowledge that the problem-solving knowledge is located in the human mind and – not in machines. A strategic aim to find solutions for data intensive problems could lay in the combination of two areas, which bring ideal pre-conditions: Human–Computer Interaction (HCI) and Knowledge Discovery (KDD). HCI deals with questions of human perception, cognition, intelligence, decision-making and interactive techniques of visualization, so it centers mainly on supervised methods. KDD deals mainly with questions of machine intelligence and data mining, in particular with the development of scalable algorithms for finding previously unknown relationships in data, thus centers on automatic computational methods. A proverb attributed perhaps incorrectly to Albert Einstein illustrates this perfectly: "Computers are incredibly fast, accurate, but stupid. Humans are incredibly slow, inaccurate, but brilliant. Together they may be powerful beyond imagination". Consequently, a novel approach is to combine HCI & KDD in order to enhance human intelligence by computational intelligence.

Keywords: Human–Computer Interaction (HCI), Knowledge Discovery in Data (KDD), HCI-KDD, E-Science, Interdisciplinary, Intersection science.

1 Challenges in the Data Intensive Sciences

Through the continuing exponential growth of data size and complexity, along with increasing computational power and available computing technologies, the data

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 319–328, 2013.

intensive sciences gain increasing importance [1]. E-Science is being advanced as a new science along side theoretical science, experimental science, and computational science, as a fundamental research paradigm [2]. Meanwhile, it is established as the *fourth paradigm* in the investigation of nature, after theory, empiricism, and computation [3], [4].

One of the grand challenges in our networked 21st century is the large, complex, and often weakly structured [5], [6], or even unstructured data [7]. This increasingly large amount of data requires new, efficient and user-friendly solutions for handling the data. With the growing expectations of end-users, traditional approaches for data interpretation often cannot keep pace with demand, so there is the risk of delivering unsatisfactory results. Consequently, to cope with this rising flood of data, new computational and user-centered approaches are vital.

Let us look, for example, at the life sciences: biomedical data models are characterized by significant complexity [8], [9], making manual analysis by the end users often impossible [10]. At the same time, experts are able to solve complicated problems almost intuitively [11], often enabling medical doctors to make diagnoses with high precision, without being able to describe the exact rules or processes used during their diagnosis, analysis and problem solving [12]. Human thinking is basically a matter of the "plasticity" of the elements of the nervous system, whilst our digital computers (Von-Neuman machines) do not have such "plastic" elements [13] and according to Peter Naur for understanding human thinking we need a different, non-digital approach, one example given by his Synapse-State theory [14].

Interestingly, many powerful computational tools advancing in recent years have been developed by separate communities with different philosophies: Data mining and machine learning researchers tend to believe in the power of their statistical methods to identify relevant patterns – mostly automatic, without human intervention, however, the dangers of modeling artifacts grow when end user comprehension and control are diminished [15], [16], [17], [18]. Additionally, mobile, ubiquitous computing and sensors everywhere, together with low cost storage, will accelerate this avalanche of data [19], and there will be a danger of drowning in data but starving for knowledge, as Herbert Simon pointed it out 40 years ago: "A wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the over abundance of information sources that might consume it" [20].

Consequently, it is a grand challenge to work towards enabling effective human control over powerful machine intelligence by the integration of machine learning methods and visual analytics to support human insight and decision support [21], the latter is still *the* core discipline in biomedical informatics [8].

A synergistic combination of methodologies, methods and approaches of two areas offer ideal conditions for addressing these challenges: HCI, with its emphasis on human intelligence, and KDD, dealing with computational intelligence – with the goal of supporting human intelligence with machine intelligence – to discover new, previously unknown insights within the flood of data.

The main contribution of HCI-KDD is, following the notion: "science is to test ideas, engineering is to put these ideas into business" [22], to enable end users to find and recognize previously unknown and potentially useful and usable information. It may be defined as the process of identifying novel, valid and potentially useful data patterns, with the goal of understanding these data patterns for decision support.

2 Human–Computer Interaction in a Nutshell

HCI evolved from computer science interest in input-output technologies [23]; hence it is a relatively young discipline, with its beginnings in the period when computers just began to have input-output devices on perception and cognition in interacting with machines [24]. Because of multimedia applications, the importance of HCI continues to increase [25], [26], [27]. Today, the main focus of HCI research is on human issues, including perception, cognition, human intelligence and interaction with any kind of information at any type of device [28].

A recent and perfect example of HCI research is the area of visual analytics, which has established its own scientific community [29] and maybe had it origins in Shneiderman's *visual analytics mantra* ("overview first, zoom/ filter, details on demand" [30]) and Keim's *extended mantra* ("Analyse First – Show the Important – Zoom, Filter and Analyse Further – Details on Demand" [29], [31]).

Keim et al. [29] provided a formal description of the visual analytics process, following some notions of Van Wijk [32], which demonstrates a core research of HCI relevant to solve problems involved with complex data:

Let the input data sets be $S = S_1, \dots, S_m$ whereas each set $S_i, i \in (1, \dots, n)$, consists of attributes $A_{i1}, \dots, A_{ik}$.

The output of this process will be called **Insight *I*.** This I is either obtained directly from a set of created **Visualizations *V***, or through confirmation of stated **Hypotheses *H*** as the results of the automatic analysis methods.

Formally this process can be described as a transformation $F : S \rightarrow I$, whereas F is a concatenation of functions $f \in \{D_W, V_X, H_Y, U_Z,\}$ defined as follows:

D_W describes the basic pre-processing functionality with $D_W : S \rightarrow S$ and $W \in \{T, C, SL, I\}$, including the data transformation functions D_T, data cleaning functions D_C, data selection functions D_{SL}, and data integration functions D_I, which are required to make the analysis functions applicable to the data set.

V_W, $W \in \{S, H\}$ symbolizes the visualization functions, which are either functions visualizing the data $V_S: S \rightarrow V$ or functions visualizing hypotheses $V_H: H \rightarrow V$.

$H_Y, Y \in \{S, V\}$ represents the hypothesis generating process.

Important in this view is that Keim et al. distinguished between functions that generate hypotheses from data $H_S : S \rightarrow H$ and functions that generate hypotheses from visualizations $H_V : V \rightarrow H$.

Human–Computer Interactions$U_Z, Z \in \{V, H, CV, CH\}$ are thus an integral part in this formal model. Such interactions may either affect visualizations only (i.e. selecting, zooming, …) $U_V : V \rightarrow V$, or the user interactions can affect hypotheses only by generating new hypotheses on the basis of the given ones $U_H : H \rightarrow H$.

Insight can be gained from visualizations $U_{CV} : V \rightarrow I$ or from the hypothesis itself $U_{CH} : H \rightarrow I$.

This leads to a main research question in HCI: "What is interesting?" [33]. Beale (2007) [34] brings it to point: Interest is an essentially human construct, a perspective on relationships between data influenced by tasks, personal preferences and experience. Interest, like beauty, is in the eye of the beholder and is strongly dependent on previous knowledge and past experience [35]. For a correct semantic interpretation a computer would need to understand the context in which a term is

presented; however, a comprehension of complex context is beyond computation. For this and many other reasons we cannot leave the search for knowledge to computers alone. We must be able to guide them as to what it is we are looking for and on which areas to focus their phenomenal computing power. Hence, in order for a data mining system to be generically useful to us, it must have some way in which we can indicate what is interesting, and for that to be dynamic and changeable [36].

Another recent and relevant research route in HCI is Attention Routing, which is a new idea introduced by Chau et al. [37] to overcome one critical problem in visual analytics: to help end users locate good starting points for analysis. Attention is of enormous importance for human information processing and through the process of selective attention, only selected subsets of the vast collection of incoming information becomes designated for further processing [38], [39], [40].

Based on anomaly detection [41], attention routing methods quasi-channel the end-user's attention through massive data structures, e.g. network structures, towards interesting nodes or sub-graphs that do not conform to "normal" behaviour. Such abnormalities often represent new knowledge that directly leads to novel insights.

From the early beginning, HCI research was focused on the measurement of human performance, heavily borrowing from the cognitive sciences [42],[35]; today HCI-research includes a large collection of various methods [43], [44], [45].

3 Knowledge Discovery in a Nutshell

Maimon & Rokach (2010) [46] define Knowledge Discovery in Databases (KDD) as an automatic, exploratory analysis and modeling of large data repositories and the organized process of identifying valid, novel, useful and understandable patterns from large and complex data sets. Data Mining (DM) is the core of the KDD process [47].

The term KDD actually goes back to the machine learning and Artificial Intelligence (AI) community [48]. Interestingly, the first application in this area was again in medical informatics: The program Rx was the first that analyzed data from about 50,000 Stanford patients and looked for unexpected side-effects of drugs [49]. The term really became popular with the paper by Fayyad et al. (1996) [50], who described the KDD process consisting of 9 subsequent steps:

1. Learning from the application domain: includes understanding relevant previous knowledge, the goals of the application and a certain amount of domain expertise;

2. Creating a target dataset: includes selecting a dataset or focusing on a subset of variables or data samples on which discovery shall be performed;

3. Data cleansing(and preprocessing): includes removing noise or outliers, strategies for handling missing data, etc.);

4. Data reduction and projection: includes finding useful features to represent the data, dimensionality reduction, etc.;

5. Choosing the function of data mining: includes deciding the purpose and principle of the model for mining algorithms (e.g., summarization, classification, regression and clustering);

6. Choosing the data mining algorithm: includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate (e.g., models for categorical data are different from models on vectors over reals) and matching a particular data mining method with the criteria of the KDD process;

7. Data mining: searching for patterns of interest in a representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modeling, dependency and line analysis;

8. Interpretation: includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns and translating the useful ones into terms understandable by users;

9. Using discovered knowledge: includes incorporating this knowledge into the performance of the system, taking actions based on the knowledge or documenting it and reporting it to interested parties, as well as checking for, and resolving, potential conflicts with previously believed knowledge.

The repertoire of methods in KDD is enormous, ranging from classical supervised methods, for example classification trees, pruning, inducers (ID3, C4.5, CART, CHAID, QUEST, and many others), Bayesian networks, regression frameworks (Regression Splines, Smoothing Splines, Locally weighted Regression), Support Vector Machines (Hyperplane Classifiers, Non-Separable SVM Models, etc.), Rule Induction (e.g. LEMI Algorithm), to classical unsupervised methods, including Clustering Algorithms, Distance Measures, Similarity Functions, Evaluation Criteria Measures, etc., Association Rules, Frequent Set Mining, Constraint-based Data Mining (e.g. Apriori, Max-Miner, MultipleJoins, Reorder, Direct, CAP, SPIRIT, etc.), Link Analysis (e.g. Social Network Analysis), to advanced Methods including Evolutionary Algorithms, Reinforcement Learning Methods, Neural Networks, Granular Computing & Rough Sets, Swarm Intelligence, Fuzzy Logic, Graph Entropy, topological Methods etc. etc. [46], [51], [52], [53], [54]. There are sheer endless possibilities of applications to solve various problems in different areas; here only a few recent examples[55], [56],[57], [58], [59], [60]. A tool to test methods, aiming at finding the most appropriate solution for a specific problem can be found here [61].

4 The Novelty of Combining HCI-KDD

To attack the grand challenges described above, a novel and promising approach is to combine the best of two worlds – HCI-KDD with the main goal of supporting human intelligence with machine intelligence – to discover new, previously unknown insights into data [5].

One could now argue: Why do we need a combination of two fields, each independently large enough? One answer is that the combination enables the provision of mutual benefits for both disciplines and allows problems to be tackled when they cannot be solved with a single disciplinary view. Likewise, cross-disciplinary or

inter-disciplinary views fall short: whilst cross-disciplinary approaches [22] are basically researching outside the scope of their component disciplines, without tight cooperation or integration from other relevant disciplines inter-disciplinary approaches are a "mix of disciplines" [62]. Consequently, trans-disciplinary approaches tend towards a beneficial fusion of disciplines, including reaching a business impact: team members from various disciplines work together on a shared problem by the application of shared conceptual frameworks, which draw together concepts, theories and approaches from those disciplines [63], [64], [65].

Such trans-disciplinary research is carried out with the explicit intent to solve multidimensional, complex problems, particularly problems (such as those related to sustainability) that involve an interface of human and natural systems [64].

This brings us back to the proverb [66] mentioned at the end of the abstract:

> Computers are incredibly fast, accurate, and stupid.
> Human beings are incredibly slow, inaccurate, and brilliant.
> Together they are powerful beyond imagination.

A strategic, synergistic and consequent combination of aspects from HCI and KDD exactly addresses this proverb.

5 Conclusion

Let us summarize the essence of the two fields:

Human–Computer Interaction (HCI), deals mainly with aspects of human perception, cognition, intelligence, sense-making and most of all the interaction between human and machine.
Knowledge Discovery from Data (KDD), deals mainly with aspects of machine intelligence, and in particular with the development of algorithms for automatic data mining.

Both disciplines have large portions of deep, unexplored and complementary subfields. A large challenge, and a possible solution to many current problems in the data-intensive sciences, may be found at the intersection of HCI and KDD.

Consequently, a novel approach is to combine HCI & KDD [5] in order to enhance human intelligence by computational intelligence. The main contribution of HCI-KDD is to *enable* end users to *find and recognize* previously unknown and potentially useful and usable information. It may be defined as the process of identifying novel, valid and potentially useful data patterns, with the goal of *understanding these data patterns* [67].

To visualize this, let us look at Figure 1: The domain expert possesses explicit domain knowledge, and by enabling him to interactively look at data sets he may be able to identify, extract and understand useful information in order to gain new, previously unknown knowledge and insight into his data sets [56].

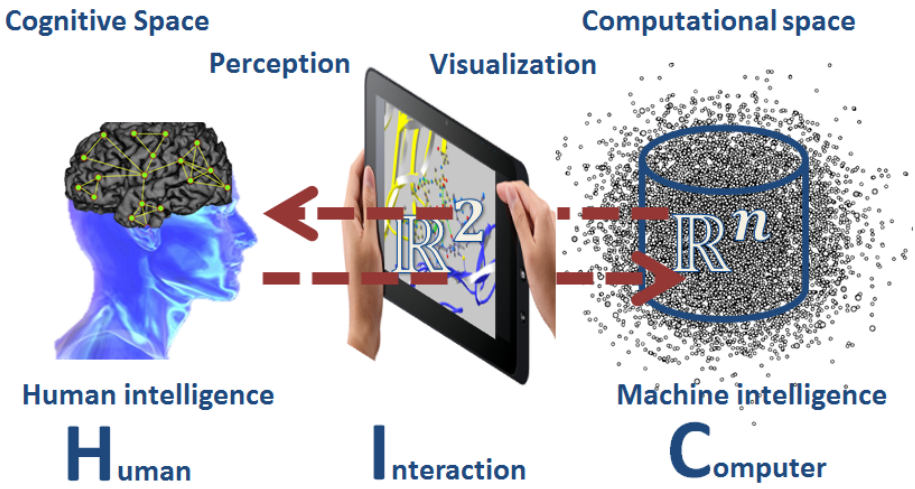


Fig. 1. We assume that "thinking" and the formation of hypotheses takes place within the human brain, enabling the interactive analysis of information properties through a synergetic combination of research on human intelligence and computational intelligence [5]

Acknowledgements. I thank the hci4all.at team and all expert members of the task force HCI-KDD.

References

1. Kouzes, R.T., Anderson, G.A., Elbert, S.T., Gorton, I., Gracio, D.K.: The changing paradigm of data-intensive computing. Computer 42, 26–34 (2009)
2. Hey, T., Gannon, D., Pinkelman, J.: The Future of Data-Intensive Science. Computer 45, 81–82 (2012)
3. Bell, G., Hey, T., Szalay, A.: Beyond the data deluge. Science 323, 1297–1298 (2009)
4. Buxton, B., Hayward, V., Pearson, I., Kärkkäinen, L., Greiner, H., Dyson, E., Ito, J., Chung, A., Kelly, K., Schillace, S.: Big data: the next Google. Interview by Duncan Graham-Rowe. Nature 455, 8 (2008)
5. Holzinger, A.: On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human–Computer Interaction & Biomedical Informatics. In: DATA 2012, pp. IS9–IS20. INSTICC, Rome (2012)
6. Holzinger, A.: Weakly Structured Data in Health-Informatics: The Challenge for Human-Computer Interaction. In: Baghaei, N., Baxter, G., Dow, L., Kimani, S. (eds.) Proceedings of INTERACT 2011 Workshop: Promoting and Supporting Healthy Living by Design, Lisbon, Portugal. IFIP, pp. 5–7 (2011)
7. Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A., Hofmann-Wellenhof, R.: Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an assistive technology in the biomedical field. In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013. LNCS, vol. 7947, pp. 13–24. Springer, Heidelberg (2013)
8. Holzinger, A.: Biomedical Informatics: Computational Sciences meets Life Sciences. BoD, Norderstedt (2012)

9. Akil, H., Martone, M.E., Van Essen, D.C.: Challenges and opportunities in mining neuroscience data. Science 331, 708–712 (2011)
10. Dugas, M., Schmidt, K.: Medizinische Informatik und Bioinformatik. Springer, Heidelberg (2003)
11. Polanyi, M.: Personal Knowledge: Towards a Post-Critical Philosophy. Nature Publishing Group (1974)
12. Popper, K.R.: Alles Leben ist Problemlösen. Piper, München (1996)
13. Naur, P.: Computing versus human thinking. Communications of the ACM 50, 85–94 (2007)
14. Naur, P.: The neural embodiment of mental life by the synapse-state theory. Naur. Com Publishing (2008)
15. Shneiderman, B.: Inventing Discovery Tools: Combining Information Visualization with Data Mining. In: Jantke, K.P., Shinohara, A. (eds.) DS 2001. LNCS (LNAI), vol. 2226, pp. 17–28. Springer, Heidelberg (2001)
16. Shneiderman, B.: Inventing Discovery Tools: Combining Information Visualization with Data Mining. Information Visualization 1, 5–12 (2002)
17. Shneiderman, B.: Creativity support tools. Communications of the ACM 45, 116–120 (2002)
18. Shneiderman, B.: Creativity support tools: accelerating discovery and innovation. Communications of the ACM 50, 20–32 (2007)
19. Butler, D.: 2020 computing: Everything, everywhere. Nature 440, 402–405 (2006)
20. Simon, H.A.: Designing Organizations for an Information-Rich World. In: Greenberger, M. (ed.) Computers, Communication, and the Public Interest, pp. 37–72. The Johns Hopkins Press, Baltimore (1971)
21. Holzinger, A.: Interacting with Information: Challenges in Human-Computer Interaction and Information Retrieval (HCI-IR). In: IADIS Multiconference on Computer Science and Information Systems (MCCSIS), Interfaces and Human-Computer Interaction, pp. 13–17. IADIS, Rome (2011)
22. Holzinger, A.: Successful Management of Research and Development. BoD, Norderstedt (2011)
23. Von Neumann, J.: The Computer and the Brain. Yale University Press, New Haven (1958)
24. Card, S.K., Moran, T.P., Newell, A.: The psychology of Human-Computer Interaction. Erlbaum, Hillsdale (1983)
25. Helander, M. (ed.): Handbook of Human-Computer Interaction. North Holland, Amsterdam (1990)
26. Holzinger, A.: Multimedia Basics. Learning. Cognitive Basics of Multimedia Information Systems, vol. 2. Laxmi-Publications, New Delhi (2002)
27. Ebert, A., Gershon, N., Veer, G.: Human-Computer Interaction. Künstl. Intell. 26, 121–126 (2012)
28. Hooper, C.J., Dix, A.: Web science and human-computer interaction: forming a mutually supportive relationship. Interactions 20, 52–57 (2013)
29. Keim, D., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual Analytics: Scope and Challenges. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 76–90. Springer, Heidelberg (2008)
30. Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages, pp. 336–343 (1996)
31. Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F. (eds.): Mastering the Information Age: Solving Problems with Visual Analytics. Eurographics, Goslar (2010)

32. Van Wijk, J.J.: The value of visualization. In: Visualization, VIS 2005, pp. 79–86. IEEE (2005)
33. Dervin, B.: Sense-making theory and practice: an overview of user interests in knowledge seeking and use. J. Knowl. Manag. 2, 36–46 (1998)
34. Beale, R.: Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and Web browsing. International Journal of Human-Computer Studies 65, 421–433 (2007)
35. Holzinger, A., Kickmeier-Rust, M., Albert, D.: Dynamic Media in Computer Science Education; Content Complexity and Learning Performance: Is Less More? Educational Technology & Society 11, 279–290 (2008)
36. Ceglar, A., Roddick, J., Calder, P.: Chapter 4: Guiding Knowledge Discovery through Interactive Data Mining. In: Pendharkar, P. (ed.) Managing Data Mining Technologies in Organizations: Techniques and Applications, pp. 45–86. Idea Group Publishing, Hershey (2003)
37. Chau, D.H., Myers, B., Faulring, A.: What to do when search fails: finding information by association. In: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 999–1008. ACM, Florence (2008)
38. Shiffrin, R.M., Gardner, G.T.: Visual Processing Capacity and Attention Control. Journal of Experimental Psychology 93, 72 (1972)
39. Kahneman, D.: Attention and Effort. Prentice-Hall, Englewood Cliffs (1973)
40. Duncan, J.: Selective attention and the organization of visual information. Journal of Experimental Psychology: General 113, 501–517 (1984)
41. Chandola, V., Banerjee, A., Kumar, V.: Anomaly Detection: A Survey. ACM Computing Surveys 41 (2009)
42. Holzinger, A., Kickmeier-Rust, M.D., Wassertheurer, S., Hessinger, M.: Learning performance with interactive simulations in medical education: Lessons learned from results of learning complex physiological models with the HAEMOdynamics SIMulator. Computers & Education 52, 292–301 (2009)
43. Lazar, J., Feng, J.H., Hochheiser, H.: Research Methods in Human-Computer Interaction. Wiley, Chichester (2010)
44. Cairns, P., Cox, A.L. (eds.): Research Methods for Human-Computer Interaction. Cambridge University Press, Cambridge (2008)
45. Nestor, P.G., Schutt, R.K.: Research Methods in Psychology: Investigating Human Behavior. Sage Publications (2011)
46. Maimon, O., Rokach, L. (eds.): Data Mining and Knowledge Discovery Handbook, 2nd edn. Springer, Heidelberg (2010)
47. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2011)
48. Piatetsky-Shapiro, G.: Knowledge discovery in databases: 10 years after. ACM SIGKDD Explorations Newsletter 1, 59–61 (2000)
49. Blum, R.L., Wiederhold, G.C.: Studying hypotheses on a time-oriented clinical database: an overview of the RX project. In: Computer-Assisted Medical Decision Making, pp. 245–253. Springer (1985)
50. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39, 27–34 (1996)
51. Piateski, G., Frawley, W.: Knowledge discovery in databases. MIT Press, Cambridge (1991)
52. Cios, J., Pedrycz, W., Swiniarski, R.: Data Mining in Knowledge Discovery. Academic Publishers (1998)

53. Liu, H., Motoda, H.: Feature selection for knowledge discovery and data mining. Springer, Heidelberg (1998)
54. Fayyad, U.M., Wierse, A., Grinstein, G.G.: Information visualization in data mining and knowledge discovery. Morgan Kaufmann Pub. (2002)
55. Billinger, M., Brunner, C., Scherer, R., Holzinger, A., Müller-Putz, G.: Towards a framework based on single trial connectivity for enhancing knowledge discovery in BCI. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 658–667. Springer, Heidelberg (2012)
56. Holzinger, A., Scherer, R., Seeber, M., Wagner, J., Müller-Putz, G.: Computational Sensemaking on Examples of Knowledge Discovery from Neuroscience Data: Towards Enhancing Stroke Rehabilitation. In: Böhm, C., Khuri, S., Lhotská, L., Renda, M.E. (eds.) ITBAM 2012. LNCS, vol. 7451, pp. 166–168. Springer, Heidelberg (2012)
57. Holzinger, A., Stocker, C., Peischl, B., Simonic, K.-M.: On Using Entropy for Enhancing Handwriting Preprocessing. Entropy 14, 2324–2350 (2012)
58. Holzinger, A., Stocker, C., Bruschi, M., Auinger, A., Silva, H., Gamboa, H., Fred, A.: On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 646–657. Springer, Heidelberg (2012)
59. Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Winkler, S.M., Schaller, S., Holzinger, A.: On text preprocessing for opinion mining outside of laboratory environments. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 618–629. Springer, Heidelberg (2012)
60. Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., Holzinger, A.: Opinion Mining on the Web 2.0 – Characteristics of User Generated Content and Their Impacts. In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013. LNCS, vol. 7947, pp. 35–46. Springer, Heidelberg (2013)
61. Holzinger, A., Zupan, M.: KNODWAT: A scientific framework application for testing knowledge discovery methods for the biomedical domain. BMC Bioinformatics 14, 191 (2013)
62. Holzinger, A.: Process Guide for Students for Interdisciplinary Work in Computer Science/Informatics, 2nd edn. BoD, Norderstedt (2010)
63. Mobjörk, M.: Consulting versus participatory transdisciplinarity: A refined classification of transdisciplinary research. Futures 42, 866–873 (2010)
64. Wickson, F., Carew, A.L., Russell, A.W.: Transdisciplinary research: characteristics, quandaries and quality. Futures 38, 1046–1059 (2006)
65. Lawrence, R.J., Després, C.: Futures of Transdisciplinarity. Futures 36, 397–405 (2004)
66. http://www.benshoemate.com/2008/11/30/einstein-never-said-that/
67. Funk, P., Xiong, N.: Case-based reasoning and knowledge discovery in medical applications with time series. Comput. Intell. 22, 238–253 (2006)

Making Sense of Open Data Statistics with Information from Wikipedia

Daniel Hienert, Dennis Wegener, and Siegfried Schomisch

GESIS – Leibniz Institute for the Social Sciences,
Unter Sachsenhausen 6-8, 50667 Cologne, Germany
{Daniel.Hienert,Dennis.Wegener,Siegfried.Schomisch}@gesis.org

Abstract. Today, more and more open data statistics are published by governments, statistical offices and organizations like the United Nations, The World Bank or Eurostat. This data is freely available and can be consumed by end users in interactive visualizations. However, additional information is needed to enable laymen to interpret these statistics in order to make sense of the raw data. In this paper, we present an approach to combine open data statistics with historical events. In a user interface we have integrated interactive visualizations of open data statistics with a timeline of thematically appropriate historical events from Wikipedia. This can help users to explore statistical data in several views and to get related events for certain trends in the timeline. Events include links to Wikipedia articles, where details can be found and the search process can be continued. We have conducted a user study to evaluate if users can use the interface intuitively, if relations between trends in statistics and historical events can be found and if users like this approach for their exploration process.

1 Introduction

Nowadays, a mass of open data statistics is available from providers like Eurostat, the United Nations or The World Bank. Eurostat for example offers about 4,500 different statistics with a wide topical range from General and Regional Statistics, Economics and Finance, Population and Social Conditions, Industry, Trade and Services, Agriculture, Forestry and Fisheries, International Trade, Transportation, Environment and Energy or Science and Technology.

Visualization software on the Web, such as Gapminder [22], makes use of these statistics and presents them in interactive graphics like diagrams or maps. Users, from laymen to politicians, can explore the data and find trends and correlations in order to "easily improve their understanding about the complex society"[1].

However, statistics - even interactively visualized - are not self-explanatory. They reflect trends for a statistical indicator (*what*) in a certain time period (*when*) and for various countries (*where*). But, these trends often base on or are related to certain events in the topic of the statistic. From a user's perspective, one can ask *why* there is

[1] http://www.gapminder.org/faq_frequently_asked_questions/

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 329–344, 2013.

a certain trend. Which influential factors lead to the trend shown in the statistic or may be related? Additional information, topically related to statistics, can help users to interpret them, get background information and hints where to continue in the search process.

We have created a prototype user interface that gives access to thousands of open data statistics from The World Bank, Eurostat and Gapminder. Users can interactively explore these statistics in different diagrams and map views. Additionally, for the Gapminder statistics, we aim at finding topically related historical events to present them in a timeline.

However, there is not much meta-information about the statistic except the title or sometimes a category. Using only keywords from the title leads to very few results when querying related information like historical events. Therefore, we use a query expansion method based on the Wikipedia and the DBpedia corpus to expand a search query with additional related terms based on the statistic title. Statistical visualizations and a timeline with the related events found are then interactively connected in a user interface to let the user explore relationships and explanations for trends in the statistic. Groups of events found with query expansion can be hidden or revealed with facets. In a user test we evaluate if users are able to use the prototype for the exploration of statistical data and related information.

Section 2 gives an overview of related work. Section 3 presents our approach for the combination of open data statistics and topically related historical events. Section 4 presents a user study to examine the exploratory search process. We discuss the results and conclude in Section 5.

2 Related Work

A lot of different data providers give access to open data statistics with focus on the world development, e.g. the United Nations, Eurostat or The World Bank. Most datasets contain information on a statistical indicator (often with several dimensions) for countries over time. Gapminder for example aggregates data from several providers and offers about 500 different indicators for about 200 countries and with a temporal coverage beginning from 1800 AD till today. Some of these data providers have recently also developed web-based visualization tools for their statistical data. Most of these web applications use maps and line/bar charts to show indicator development over time and for different countries. Gapminder uses a different approach and shows two related indicators in a scatterplot. Similar to statistical software for scientists, such as SPSS, STATA or R, one can see correlations between these indicators. Gapminder additionally uses an animation to show the development of correlation over time for all countries of the world.

For the semantic modeling of events in RDF a number of ontologies is available, e.g. EVENT[2], LODE [24], SEM [8], EventsML[3] and F [23]. A comparison of such ontologies can be found in [24]. Furthermore, there exist ontologies and systems for the annotation of events in the timeline. Gao & Hunter [7] make use of ontologies not only for the semantic markup of events, but also for the modeling of timelines,

[2] http://motools.sourceforge.net/event/event.html
[3] http://www.iptc.org/site/News_Exchange_Formats/EventsML-G2/

relationships between events and for the annotation of events of different timelines. Therefore, not only events can be referenced in the Linked Data Cloud, but also relations and annotations. In a web-based interface, users can search for geological events with a high impact such as earthquakes, tsunamis and volcanic eruptions, can analyze relationships and make connections between them with annotations.

The simple representation of events in timelines can be implemented with various web-based tools such as SIMILE[4] widgets. The use of timelines can cause user interaction difficulties, for example, if too many events are visualized and the user loses the navigational overview [16]. Kumar et al. [15] propose a theoretical model for timelines on the storage and presentation layer with levels like content, operations, browsing and editing. An example implementation of this model allows the creation, visualization and browsing of bibliographic metadata. LifeLines [21] utilizes the arrangement of several timelines in one view. Facets of personal records, either flat or hierarchic, can individually be selected by buttons or trees to give an overview or discover relationships. Other complex timeline visualizations like Semtime [14] or Sematime [26] display in a similar fashion several stacked timelines with additionally depicted time-dependent semantic relations between events. Advanced interaction techniques include hierarchical filtering and navigation like zooming or expanding sub-timelines. Sense.us [9] is a web-based visualization system with an emphasis on the social and collaborative aspect. Users can share visualizations, e.g. a chart of US census data, and can discuss and comment on different states of the visualization. Bookmarking ability of different states and graphical annotation tools allow annotating and discussing certain data points or trends in statistical visualizations in order to make sense of the pure data. ChronoViz [6] is a system that uses the timeline metaphor for the simultaneous display of multiple data sources such as video, audio, logs, sensor data and transcriptions. Also here, users can add annotations with a digital pen and use them as anchor links. Aigner et al. [2] give a systematic overview of visualizing time-oriented data based on categorization criteria on the time, data and representation level. They emphasize open problems and future work like multiple views of time-oriented data and their coordination. In this sense, some case studies exist which use Multiple Coordinated Views [28] of temporal and other visualizations in domains like climate data [25] or medicine [1]. Timelines are linked to other views, so that selecting data in a timeline highlights data in other views or vice versa.

The search for relationships between pieces of information in different representations can be described by the model of Exploratory Search [17]. Multiple iterations with cognitive processing and interpretation of objects over various media like graphs, maps, text and video is needed. The user has to spend time "scanning/viewing, comparing and making qualitative judgments" for the result of "knowledge acquisition, comprehension of concepts or skills, interpretation of ideas, and comparisons or aggregation of data and concepts" [17]. Also, timelines can be part of an exploratory search process, for example, to use the time component to explore research articles in a timeline view [3].

Query expansion is used to expand a search query with additional terms like synonyms, related terms, or methods like stemming or spelling correction. For very short queries this can increase the overall recall of a search, and more relevant

[4] http://www.simile-widgets.org/timeline/

information objects can be found [27]. Query expansion based on thesauri, ontologies, co-occurrence analysis or other knowledge sources have been utilized in digital libraries of different domains like e.g. the social sciences [18] or the medical domain [13]. Also, the Wikipedia corpus has been used as a database for this purpose [19].

3 System Prototype

We have created a research prototype[5] for the visualization of open data statistics and related events. In the following, we will describe (1) the visualization of statistical data in several views, (2) the retrieval and visualization of related historical events.

3.1 Visualization of Statistical Data

The web application is implemented in PHP and the visualization and interaction component is realized with HTML5, the canvas element and JavaScript (based on a previous prototype presented in [12]). We have integrated statistical data from Eurostat, The World Bank and Gapminder. The Eurostat dataset contains 4,545 indicators, The World Bank 3,566 and Gapminder 498.

Because the prototype already contains about 8,500 statistics from different providers, it is difficult to provide an overall hierarchical categorization. Therefore, the web application provides a query interface for searching statistics by title. The user can enter keywords and an autocomplete function initially suggests matching statistics. By clicking on the link the indicator is chosen and instantly visualized.

Similar to existing solutions, the statistical data is displayed in various graphical views: a map, a bar chart and a line chart (see Figure 1).

On a Google map, indicator values for all different countries and a given year are visualized as circles of different sizes and colors at the country location. The values of the indicators are visually encoded in two dimensions: (1) the higher the indicator value, the greater the circle radius, (2) the higher the indicator value, the more the color shade goes into the red spectrum, low indicators values are encoded with bluish tones. This way, the user can instantly see where high values are clustering. For example, for the indicator *Fertility*, one can see at first glance that in 2011 still more children per woman were born in Africa than in Europe. The map has standard interaction facilities like zooming and panning, so the user can browse to the region of interest.

The bar chart shows indicator values for each country and a given year in descending order. Countries with high values are at the top, countries with low values are at the bottom of the list. As 200 countries are listed, the user can scroll the list. The width of the horizontal bars allows seeing differences between countries instantly, so the users can compare if there are only small or big differences between two countries in the list. Similar to the map, the bars are color-coded, so one can see instantly if countries have higher or lower values than the Median.

[5] `http://opendatastatistics.org`

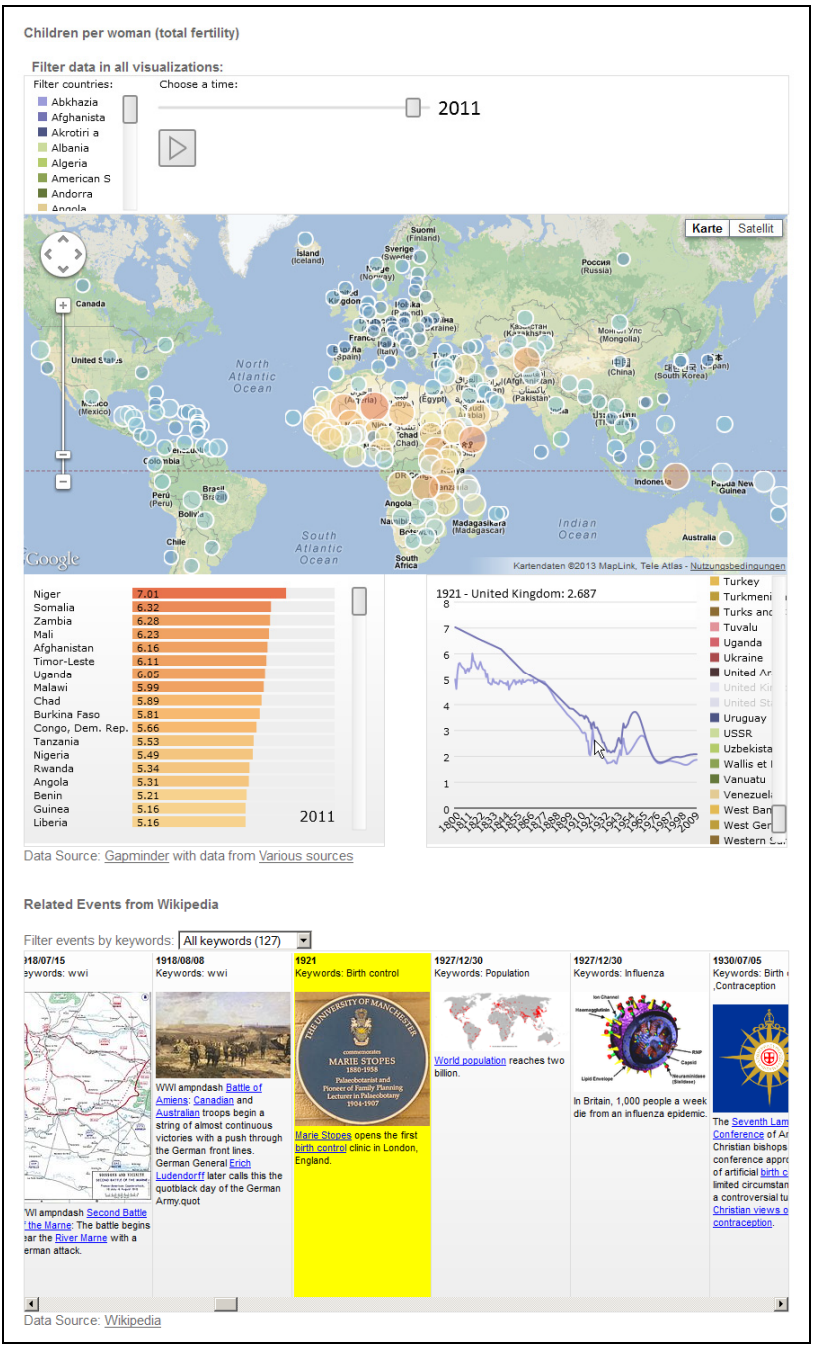


Fig. 1. Prototype with statistical visualizations and a timeline with related events for the statistic "Children per woman (Total fertility)" from Gapminder. Hovering the mouse over the data point United Kingdom – 1921 scrolls to and highlights events for the same year in the timeline.

Next to the bar chart, there is a line chart that shows the distribution of indicator values over time. Users can select a country from the chart legend or from the overall filters which results in the line showing up in the graph. Hovering with the mouse over a data point shows time, country and value at the top of the chart. Lines are color-coded similarly in the legend and in the graph.

At the top of the page, there exists a control panel for filtering all views simultaneously by one or several countries and a year. From a list, users can select individual countries which are color-coded according to countries in the line chart. In addition, there exists a slider for selecting the year. This way, the selected year for the map and bar chart view can be chosen and the user can browse through the time. With a play button all views can be animated to show data for the actual year.

3.2 Retrieval of Related Historical Events

A goal in this paper is to retrieve and display historical events that are related to a statistical indicator and therefore, link the following two datasets:

1. The Gapminder dataset includes 498 different statistical indicators. This dataset is maintained regularly, topically well chosen, includes many data sources and covers all world countries with a wide temporal coverage from 1800 till today.
2. The historical events dataset is based on a database that holds a large collection of historical events extracted from Wikipedia [10] and can be queried via an API[6]. It has been shown that machine learning with features from DBpedia is a feasible way to achieve an automatic classification of the extracted events into categories [11]. The outcome of the effort was a dataset with about 190,000 historical events covering different languages and including category information. The data subset used in this paper contains 37,859 yearly English events from 300 BC to 2013. An important requirement is that both datasets have a focus on years as a temporal reference unit. Statistical indicators are resolved for years; historical events are chosen by the Wikipedia community as an important fact for a year's history.

The statistic title or its category only provide weak information for the retrieval of related events. For example, querying the historical database for the statistic "Children per woman (total fertility)" with the keyword "fertility" leads to 0 results. The query has to be expanded with strongly related concepts to find more events and to get a higher recall. For "fertility", concepts like "contraception", "birth control", or "syphilis" must be found to retrieve historical events that have an explanatory character for trends in the statistic. The two data sources are connected in the following way (see Figure 2):

Step 1: For each of the Gapminder statistics, the title is preprocessed by removing units and stop words/characters.

Step 2: Based on the preprocessed title, the Wikipedia API is queried for an article page with a matching concept. We inspected the top 10 search results and manually selected the semantically best fitting page for each statistic. As a result, we got a

[6] `http://www.vizgr.org/historical-events/`

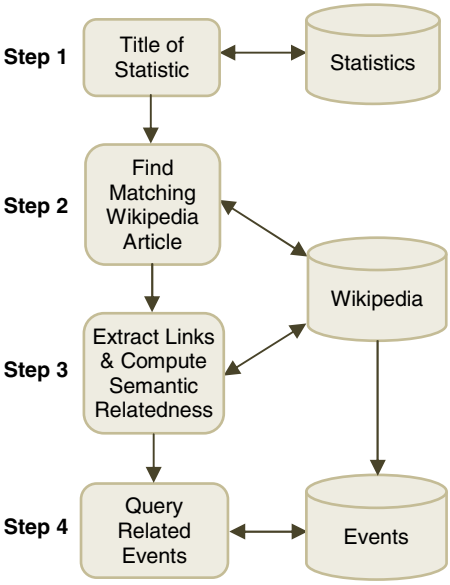


Fig. 2. Process steps to retrieve related events for a statistic

mapping to a chosen Wikipedia article for each statistical indicator. All in all, this includes 144 distinct articles, as some of the statistics belong to the same Wikipedia concept (e.g. to the Wikipedia article "Development aid"). Because statistic titles from Gapminder are really short, we decided to manually select the Wikipedia article to guarantee a well-chosen concept with balanced properties of abstractness level, size, number of links etc.

Step 3: As a next step, we want to find related concepts for the mapped concepts. Therefore, we have implemented a web service[7] that returns related concepts for an input term. The service queries in/outlinks from Wikipedia via the Web API as well as broader/narrower terms and category information from DBpedia [4] via SPARQL endpoint. For each concept, the semantic relatedness (SR) to the original term is computed. For doing so, we use the Normalized Google Distance (NGD) formula [5], but instead of taking hit counts from a search engine we use hit counts from the Wikipedia full text search.

Semantic relatedness is then computed with the following formula:

$$SR = \frac{log_{10}\big(\max(A,B)\big) - log_{10}(A \cup B)}{log_{10}(W) - log_{10}\big(\min(A,B)\big)}$$

A: Number of full text search hits in Wikipedia for concept one

B: Number of full text search hits in Wikipedia for concept two

$A \cup B$: Number of full text search hits in Wikipedia for concept one AND concept two.

W: Number of articles in Wikipedia.

[7] http://www.vizgr.org/relatedterms/

In a separate evaluation with evaluation datasets included in the sim-eval framework [20] we found that this approach achieved a Spearman correlation up to 0.729 for human judged datasets and P(20) up to 0.934 for semantic relation datasets. We use all concepts with a SR > 0.3 as keywords for the following step, which filters very broad or non-related concepts. This way, we could compute 77,202 related concepts for 498 statistics with on average 155 keywords per statistic.

Step 4: In the last step, we query events based on the concepts found in the previous step. In detail, we query the Historical Events-API for all events that include at least one of these keywords in the event description and lies within the time interval of the statistic. For all 498 statistics, querying only the Wikipedia equivalent for a statistic returns in sum 8,450 events, which means on average about 17 events per statistic. Using the presented approach for query expansion, for all statistics in sum 137,921 (not distinct) events are returned, which means on average about 279 per statistic. Based on the returned events a timeline is build. To illustrate the different steps, we present two examples related to the topics *fertility* and *earthquakes*.

Example *"Fertility"*:
The title of the statistic from Gapminder in the first example is "Children per woman (total fertility)". The title was mapped to the Wikipedia concept "Fertility". Querying related concepts returns 61 keywords with an SR value higher than 0.3 (compare Table 1). These keywords were then used to query the Historical Events-API which results in 22 matching events.

Example *"Earthquake"*:
The title of the statistical indicator in the second example is "Earthquake - affected annual number". For this statistic, the Wikipedia article "Earthquake" was selected. 132 related concepts were returned with an SR value higher than 0.3. For these keywords the Historical Events-API returned 1,008 events.

3.3 Visualizing Related Historical Events

Below the area that visualizes statistical data, a scrollable timeline shows queried historical events in a timeline (see Figure 1, at the bottom). The individual events consist of a date, keywords, a description and, if available, a thumbnail. Links in the event description allow browsing to Wikipedia articles and to read details there. Events for a certain keyword can be filtered with a select box above the timeline. The line chart that shows the indicator distribution over time for certain countries and the timeline are linked by a brushing-and-linking mechanism. When a user brushes over a data point in the line chart, the timeline automatically scrolls to events from the same year and highlights them with a yellow background or vice versa.

Table 1. Top 20 Wikipedia concepts including values for semantic relatedness for "fertility" and "earthquake"

Links for Fertility	SR	Links for Earthquake	SR
Fertility rate	0.714	Tsunami	0.693
Total fertility rate	0.657	Seismic	0.651
Infertility	0.638	2011 Tohoku earthquake and tsunami	0.584
Sperm	0.622	Seismology	0.581
Crude birth rate	0.590	Epicenter	0.572
Contraception	0.588	Aftershock	0.566
Fertile	0.577	Disaster	0.554
Fertilization	0.576	2004 Indian Ocean earthquake	0.549
Gestation	0.572	2010 Haiti earthquake	0.543
Uterus	0.571	1906 San Francisco earthquake	0.529
Fertility clinic	0.561	Northridge earthquake	0.523
Ovary	0.559	Volcano	0.521
Menstrual cycle	0.558	Megathrust earthquake	0.518
Menopause	0.552	Subduction	0.512
Pregnancy	0.544	Seismologist	0.510
Endocrinology	0.535	Natural disaster	0.507
Fetus	0.535	Landslide	0.502
Fecundity	0.533	Foreshock	0.496
IUD	0.513	Mercalli intensity earthquake	0.491
Sperm count	0.512	1755 Lisbon earthquake	0.489

4 User Study

We have conducted a user study to examine the following research questions:

- Can participants use the interface intuitively?
- How do users normally search for additional information of trends in a statistical indicator?
- Can our approach of showing related historical events in a timeline help users to find interesting background information and starting points for further exploration?
- Which advantages and limitations does our approach have?

4.1 Method and Participants

The participants were asked to carry out a set of tasks in the prototype and to fill out a questionnaire after each task with their actions performed, results found, time required, experienced difficulty level and comments. First, the users could familiarize themselves with the environment for 5 minutes. After that, they had to complete three different tasks. Finally, the users were asked to evaluate the pros and cons of our approach and to assess the overall scenario. The group of participants included eight male researchers and one female researcher, aged 27 to 40 (mean: 31 years). All had a graduate degree in computer science like Master or similar. The participants were asked to rate their experience in dealing with web-based search on a five-point-scale. The rating was 1.56 ("very good") with a standard deviation of 0.73.

4.2 Task and Questions

The participants had to handle the following tasks and answer questions in the questionnaire:

1. Fertility trends across countries
Briefly describe the development of fertility rates in the United States of America and in the United Kingdom from 1921 to 1940 as shown in the line graph (a). What are the values for each country in the years 1926 (b) and 1937 (c)?

For task 1, the users had to write down the response in a text field, answer how long the process took, assess as how difficult the task was perceived on a five-point-scale (2=very easy, 1=easy, 0=normal, -1=difficult, -2=very difficult) and were asked to give comments and suggestions.

2. Causes for the decrease in fertility after 1920 and after 1960 in the US and in the UK
(a) Find possible reasons for the decrease of fertility in the United States of America and in the United Kingdom from 1920 to 1940. (b) Find possible reasons for the decrease of fertility in the United States of America and in the United Kingdom in 1960. Try to find information outside the given prototype using other sources, e.g. Google. Do not spend more than 5 minutes for each subtask (a) and (b).

For task 2, the users had to record the search steps, the total time, the relevant information sources and a confidence score for the search result on a five-point-scale (1=very unsure, 2=unsure, 3=normal, 4=sure, 5=very sure). Furthermore, similar to task 1, they had to assess the difficulty and could give comments and suggestions. For this task, the timeline has not been visible in the UI to prevent any influences on the user. For task 3, the participants could then refresh the user interface and activate the timeline with a GET-Parameter in the URL.

3. Usage of the timeline
Analog to task 2, please find potential causes or contexts for the decrease in fertility rates in these countries from 1920 to 1940 and from 1960 based on the historical events displayed in the timeline and describe them briefly.

For task 3, users had to write down the response in a text field, enter a confidence score as in task 2 and to assess the difficulty and could give comments and suggestions as in tasks 1 and 2.

4. Comparison and evaluation of the two methods from tasks 2 and 3 and overall results

Please evaluate the overall scenario:

- What are the pros and cons of both search methods for the application scenario?
- Which search strategy would you favor for the given application scenario and why would you do so?
- Was the integrated user interface including the graphical view of information of fertility statistics and historical events in a timeline helpful for answering the questions on the decrease of fertility after 1921/1960 in both countries?

The overall scenario could be rated with a five-point-scale (2=very helpful, 1=helpful, 0=normal, -1=not helpful, -2=not helpful at all) and we left room for general comments, suggestions and criticisms.

4.3 Results

After the participants had made themselves familiar with the user interface they could all (n=9) solve task 1 successfully. The average time exposure for sub task (a) resulted in 74 seconds, while answers for the country specific values (b) and (c) took on average 16 seconds (compare Table 2). Participants rated the difficulty level of the task on average with "easy" (1.00). Task 1 showed that the user interface could be easily adopted by users for filtering and read-off processes without further explanation. Nevertheless, some points of criticism and improvement were given in terms of better scrolling functionality in the country list, more precise handling of the sliders, better highlighting of the selected countries, more distinguishable colors and zooming functionality in the line chart.

For task 2, the participants recorded 30 query steps in the questionnaire to find possible reasons for the decrease of fertility in the US/UK between 1920 and 1940 (a) and from 1960 (b). The majority of users used Google (23 times), followed by Wikipedia (6 times). Solving task 2 took 122s/118s per identified reason. Users stated Wikipedia (10 times) and other websites (18 times) as sources of information. On average, the participants had a normal confidence in the information found on the Web (3.19/3.29). The average difficulty level of task 2 was evaluated with "difficult" for (a) and "normal" for (b).

In task 3 the participants were faced with the same questions as in task 2, but now could enable and use the timeline in the user interface to find possible reasons. For the decrease in fertility from 1920 to 1940 in the USA/UK, participants named, e.g., nine times the event "*1921 - Marie Stopes opens the first birth control clinic in London, England.*", five times *World War I and II* (from several events) or two times "*1930/07/05 - The Seventh Lambeth Conference of Anglican Christian bishops opens. This conference approved the use of artificial birth control in limited circumstances, marking a controversial turning point in Christian views on contraception*." One user stated that a decrease in fertility is normal for industrial countries. For task a) the participants were confident for possible reasons ("sure", 3.67).

For the decreasing trend from 1960 the users stated, e.g., eight times "*1960/05/09 - The U.S. Food and Drug Administration announces that it will approve birth control as an additional indication for Searle's Enovid, making it the world's first approved oral contraceptive pill.*", two times "*1958/12/31 - Based on birth rates (per 1,000 population), the post-war baby boom ends in the United States as an 11-year decline in the birth rate begins (the longest on record in that country).*" or one time *Vietnam War* (from one event). For task b) the participants were confident in having found possible reasons (4.17, "sure"). The task needed 58 seconds per reason and was perceived as "normal" (0.39) difficult on average.

Table 2. Summarized Results

Task 1: Fertility trends across countries			
	Successful responses	Time needed	Difficulty level
a) Trend from 1921 to 1941	9/9	74s	„easy“ (1.00)
b)/c) Values for 1926 and 1937	9/9	16s	

Task 2: Causes for the decrease in fertility after 1920 and after 1960 in the US and in the UK					
	Query steps	Relevant sources	Confidence	Time needed per reason	Difficulty level
a) From 1920 to 1940	12x Google; 3x Wikipedia; 1x general knowledge	12x different sources; 4x Wikipedia	“normal” (3.19)	122s	“difficult” (-0.50)
b) From 1960	11x Google; 3x Wikipedia	8x different sources; 6x Wikipedia	“normal” (3.29)	118s	“normal” (-0.25)

Task 3: Usage of the timeline				
	Potential reasons	Confidence	Time needed per reason	Difficulty level
a) From 1920 to 1940	9x: Marie Stopes opended first birth control clinic in UK in 1921; 5x: World War I/II; 2x 7th Lambeth Conference; 1x: normal development for industrialized countries	“sure” (3.67)	58s	“normal” (0.39)
b) From 1960	8x: FDA - Approval and usage of oral contraceptive pill; 2x: baby boom ends; 1x: Vietnam war;	“sure” (4.17)		

For task 4, users could first compare “search on the Web” (task 2) and “search in the prototype” (task 3). The web search was advantageous to the users in terms of using familiar systems (Google, Wikipedia etc.), having access to all kinds of documents, sources and contexts, the possibility to compare the resulting information and the control of the search process. The disadvantages reported were the heavy-handed search methods to find the relevant documents or sources, the time consuming process, additional search steps, no single answers for complex questions, a lot of irrelevant information and the problem of how to translate the information need into a search query when no starting point is given.

For the search in the prototype, the participants stressed as pros that they have found the results faster and could see them directly. Furthermore, they assessed positively the single point of search as well as the integrated search environment and the synchronization between line chart and timeline. In contrast, they qualified negatively that the search quality in this system depends on the used data sources and that only Wikipedia documents were offered but no scientific studies or other information.

Seven participants would initially prefer the prototype from task 3 to get an overview over a topic and to get results quickly. Some participants stated that this information was enough; most participants explained that they would use the prototype for a quick overview and then use the web in order to get reliable and quotable data or for verification of the results of task 3.

The participants evaluated the search in the prototype tendentially as helpful, which became apparent in the average value of 0.50 ("helpful"). This was also stressed, although with constraints, in the textual evaluation. One participant thought the prototype to be convenient only for easy questions but not for complex issues in the social sciences, other users wished more data sources.

5 Discussion and Conclusion

The conducted user study gives first hints that the combination of numerical statistical data in different views and related information like events in a timeline in a single user interface can be fruitful. Participants were able to use the interface intuitively without further instructions. The results of task 2 show that participants search for related information of trends in statistics with search engines like Google and Wikipedia with a combination of what, when, and where keywords. It can be seen from the confidence values that the users were confident about having found related information. On the Web, they have access to all kinds of documents, but it was a time-consuming process and a lot of search steps were needed. Our prototype was preferred for a fast overview as a single point of search and a starting point, but, of course, did not include all sources a web search engine provides.

For the user test, we chose the very complex topic "fertility" in contrast to less complex topics like e.g. "earthquakes" (in a sense of how many other diverse complex aspects may have an influence on the indicator). Influential factors are difficult to determine for a normal web search user, because related concepts and keywords are not at hand. Table 1 gives an overview: While concepts for "fertility" are very diverse, a lot of concepts for "earthquake" include the term "earthquake" with combinations of different locations and dates. This makes it easy also for normal web search, where the query instantly leads to the matching Wikipedia article (e.g. for "2010 Haiti earthquake"). In contrast, querying for all aspects of fertility seems to be a harder process. If one is lucky, search queries like "fertility 1920s United Kingdom" (as performed by participants of our study) lead to documents were information and explanations for one location and one time period are included. However, this documents most times only provide very broad explanations for a concept and no further detailed information or links. For example, for fertility users found a

document that contains information on the development of fertility in the UK for the last century. Here, influential factors like both world wars, economic depression (late 1920s) and influenza outbreak (after World War I) are described. However, the document did not include any further information or links. Instead, users have to copy and paste keywords into a search engine. Our approach tries to take off the burden of searching relevant keywords and querying related information. In this sense, we computed related concepts and used these to query the data source.

In the presented use case, we concentrated on the combination of statistics with historical events. Since these events are carefully chosen according to their importance on a world history view by Wikipedia users, they offer a good compromise between importance, abstractness, count and temporal coverage (compare [10]). Of course, other similar data sources could be included such as news articles from the Guardian Open Platform[8] or the New York Times API[9]. However, we found that these sources provide the majority of articles for the last three decades and the query with keywords returns a mass of only rather important articles. Here, further aggregation steps have to be applied. But also other related information types can be queried with the computed concepts like Wikipedia articles, web sites, studies and surveys from which the statistics are generated, videos, images etc. Another important aspect is that the search on the Web and the search in the interface are not complementary as proposed in the user study, but users can use links in the timeline to continue with their web search.

At the end, it is always a trade-off between presenting too much or too little information, and carefully choosing important information based on only unsubstantial query information like statistic title (and related concepts), country and time.

References

1. Aigner, W., Miksch, S.: Supporting Protocol-Based Care in Medicine via Multiple Coordinated Views. In: Proceedings of the Second International Conference on Coordinated & Multiple Views in Exploratory Visualization, pp. 118–129. IEEE Computer Society, Washington, DC (2004)
2. Aigner, W., Miksch, S., Müller, W., Schumann, H., Tominski, C.: Visualizing time-oriented data-A systematic view. Comput. Graph 31(3), 401–409 (2007)
3. Alonso, O., Baeza-Yates, R., Gertz, M.: Exploratory Search Using Timelines. Presented at the SIGCHI 2007 Workshop on Exploratory Search and HCI Workshop (2007)
4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: A Nucleus for a Web of Open Data. In: Aberer, K., et al. (eds.) ISWC/ASWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
5. Cilibrasi, R.L., Vitanyi, P.M.B.: The Google Similarity Distance. IEEE Trans. Knowl. Data Eng. 19(3), 370–383 (2007)
6. Fouse, A., Weibel, N., Hutchins, E., Hollan, J.D.: ChronoViz: a system for supporting navigation of time-coded data. In: Tan, D.S., Amershi, S., Begole, B., Kellogg, W.A., Tungare, M. (eds.) CHI Extended Abstracts, pp. 299–304. ACM (2011)

[8] http://www.guardian.co.uk/open-platform
[9] http://developer.nytimes.com/

7. Gao, L., Hunter, J.: Publishing, Linking and Annotating Events via Interactive Timelines: an Earth Sciences Case Study. In: Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web, DeRiVE 2011 (2011)
8. van Hage, W.R., Malaisé, V., Segers, R.H., Hollink, L., Schreiber, G.: Design and use of the Simple Event Model (SEM). Web Semant. Sci. Serv. Agents World Wide Web 9(2), 2 (2011)
9. Heer, J., Viégas, F.B., Wattenberg, M.: Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1029–1038. ACM, New York (2007)
10. Hienert, D., Luciano, F.: Extraction of Historical Events from Wikipedia. In: Proceedings of the First International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (KNOW@LOD), Heraklion, Greece, pp. 25–36 (2012)
11. Hienert, D., Wegener, D., Paulheim, H.: Automatic Classification and Relationship Extraction for Multi-Lingual and Multi-Granular Events from Wikipedia. In: van Erp, M., van Hage, W.R., Troncy, R., Shamma, D.A. (eds.) Proceedings of the Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012), Boston, USA, pp. 1–10 (2012)
12. Hienert, D., Zapilko, B., Schaer, P., Mathiak, B.: Web-Based Multi-View Visualizations for Aggregated Statistics. In: Proceedings of the 5th International Workshop on Web APIs and Service Mashups, pp. 11:1–11:8. ACM, New York (2011)
13. Holzinger, A., Yildirim, P., Geier, M., Simonic, K.-M.: Quality-based knowledge discovery from medical text on the Web Example of computational methods in Web intelligence. In: Pasi, G., Bordogna, G., Jain, L.C. (eds.) Quality Issues in the Management of Web Information. ISRL, vol. 50, pp. 145–158. Springer, Heidelberg (2013)
14. Jensen, M.: Visualizing Complex Semantic Timelines. NewsBlib. (2003)
15. Kumar, V., Furuta, R., Allen, R.B.: Metadata visualization for digital libraries: interactive timeline editing and review. In: Proceedings of the Third ACM Conference on Digital Libraries, pp. 126–133. ACM, New York (1998)
16. Kurihara, K., Vronay, D., Igarashi, T.: Flexible timeline user interface using constraints. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1581–1584. ACM, New York (2005)
17. Marchionini, G.: Exploratory search: from finding to understanding. Commun. ACM 49(4), 41–46 (2006)
18. Mayr, P., Petras, V.: Cross-concordances: terminology mapping and its effectiveness for information retrieval. CoRR. abs/0806.3765 (2008)
19. Medelyan, O., Milne, D., Legg, C., Witten, I.H.: Mining meaning from Wikipedia. Int. J. Hum.-Comput. Stud. 67(9), 716–754 (2009)
20. Panchenko, A., Morozova, O.: A study of hybrid similarity measures for semantic relation extraction. In: Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, pp. 10–18. Association for Computational Linguistics, Stroudsburg (2012)
21. Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B.: LifeLines: visualizing personal histories. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 221–227. ACM, New York (1996)
22. Rosling, H.: Visual technology unveils the beauty of statistics and swaps policy from dissemination to access. Stat. J. Iaos J. Int. Assoc. Off. Stat. 24(1-2), 103–104 (2007)
23. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F–a model of events based on the foundational ontology dolce+DnS ultralight. In: Proceedings of the Fifth International Conference on Knowledge Capture, pp. 137–144. ACM, New York (2009)

24. Shaw, R., Troncy, R., Hardman, L.: LODE: Linking Open Descriptions of Events. In: Gómez-Pérez, A., Yu, Y., Ding, Y. (eds.) ASWC 2009. LNCS, vol. 5926, pp. 153–167. Springer, Heidelberg (2009)
25. Shimabukuro, M.H., Flores, E.F., de Oliveira, M.C.F., Levkowitz, H.: Coordinated Views to Assist Exploration of Spatio-Temporal Data: A Case Study. In: Proceedings of the Second International Conference on Coordinated & Multiple Views in Exploratory Visualization, pp. 107–117. IEEE Computer Society, Washington, DC (2004)
26. Stab, C., Nazemi, K., Fellner, D.W.: Sematime - timeline visualization of time-dependent relations and semantics. In: Bebis, G., et al. (eds.) ISVC 2010, Part III. LNCS, vol. 6455, pp. 514–523. Springer, Heidelberg (2010)
27. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 61–69. Springer-Verlag New York, Inc., New York (1994)
28. Wang Baldonado, M.Q., Woodruff, A., Kuchinsky, A.: Guidelines for using multiple views in information visualization. In: Proceedings of the Working Conference on Advanced Visual Interfaces - AVI 2000, Palermo, Italy, pp. 110–119 (2000)

Active Learning Enhanced Document Annotation for Sentiment Analysis

Peter Koncz and Ján Paralič

Dept. of Cybernetics and Artificial Intelligence,
Technical University of Košice, Slovak Republic
{peter.koncz,jan.paralic}@tuke.sk

Abstract. Sentiment analysis is a popular research area devoted to methods allowing automatic analysis of the subjectivity in textual content. Many of these methods are based on the using of machine learning and they usually depend on manually annotated training corpora. However, the creation of corpora is a time-consuming task, which leads to necessity of methods facilitating this process. Methods of active learning, aimed at the selection of the most informative examples according to the given classification task, can be utilized in order to increase the effectiveness of the annotation. Currently it is a lack of systematical research devoted to the application of active learning in the creation of corpora for sentiment analysis. Hence, the aim of this work is to survey some of the active learning strategies applicable in annotation tools used in the context of sentiment analysis. We evaluated compared strategies on the domain of product reviews. The results of experiments confirmed the increase of the corpus quality in terms of higher classification accuracy achieved on the test set for most of the evaluated strategies (more than 20% higher accuracy in comparison to the random strategy).

Keywords: sentiment analysis, active learning, semi-automatic annotation, text mining.

1 Introduction

Sentiment analysis, also called opinion mining, is a research area devoted to the methods of automatic quantification of the subjective content expressed in the form of natural language [1]. It aims to detect the presence, orientation or even the intensity of the opinion related to the object of the evaluation or its features/aspects in case of aspect-based sentiment analysis. Within the methods of sentiment analysis, one of the main streams is represented by methods based on using of machine learning algorithms, which deals with the task of sentiment analysis as with a text categorization task. However, these methods are dependent on manually annotated corpora for the training of the classifiers. Moreover these classifiers have been shown domain dependent, i.e. the classifier created for one domain is hardly portable to other domain. Hence, in real application scenario it is usually necessary to build separate corpora for different domains. In order to make the annotation process more effective methods of active learning can be utilized. However active learning is a common strategy used to

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 345–353, 2013.

increase the efficiency of classifiers creation, currently it is a lack of systematical research devoted to its application in the context of sentiment analysis. The need of the integration of active learning to annotation tools for sentiment analysis led us to the comparative evaluation of six active learning strategies. The rest of the work is divided as follows. The second chapter is devoted to related works. In the third chapter the evaluated active learning strategies are described. Consequently the fourth chapter describes their experimental evaluation. Finally, the last part is devoted to conclusions.

2 Related Work

In the last years it was published a huge amount of works devoted to sentiment analysis. Their comprehensive overview can be found in the work of Liu [2]. The methods of sentiment analysis are typically divided into two groups.

The first group is represented by so-called lexicon-based methods. These methods are usually based on sentiment dictionaries and rules for working with them [3]. Examples of sentiment dictionaries include the *MPQA subjectivity lexicon*[1], used also in our work, *Appraisal lexicon*[2]*, National Taiwan University Sentiment Dictionary*[3]or the*SentiWordNet*[4]. There are also some works devoted to the possibilities of automation of dictionaries creation, which usually try to utilize the existing structured knowledge resources. Kamps et al. [4] described a method for identification of adjectives orientation on the basis of their distance from the words *good* and *bad* in a WordNet graph. In the work of Kim and Hovy[5] another solution using WordNet was proposed, based on the extension of the set of emotional words using their synonyms and antonyms. Another solution to populate sentiment dictionaries is to use seed sets of polarity words and extend it by analysis of their coincidences with other words in the corpus of documents as in the work of Turney[6]. These dictionaries, as it will be described in the following section, can be utilized in order to compute the sentiment classification uncertainty. Unlike the above mentioned works, the method for generation of sentiment dictionaries used in this work is based only on the annotated corpora of documents.

The second group of sentiment analysis methods is represented by machine learning based or corpus based methods [3]. These methods use manually annotated corpora, on the basis of which classifiers are trained for identification of the sentiment. In the frame of this group of methods support vector machines (SVM) [1], [7–11] and naïve Bayes classifier (NBC) [10], [11] have been widely used. SVM was also shown to be the most advantageous method in the scope of our own works [1], [3]. Besides the learning algorithms feature selection methods have been also intensively studied [1], [7], [8], where information gain has been shown as one of the most effective methods. One of the main drawbacks of this group of sentiment analysis methods is their dependency on manually annotated corpora. For the robustness of the created solutions it is necessary to take into account the differences between the analyzed texts.

[1] www.cs.pitt.edu/mpqa
[2] lingcog.iit.edu/arc/appraisal_lexicon_2007b.tar.gz
[3] nlg18.csie.ntu.edu.tw:8080/lwku
[4] sentiwordnet.isti.cnr.it

Depending on the source as well as other characteristics the particular texts differ in language, size, formality or in the usage of nonlinguistic expression forms like emoticons [12]. Moreover the created classifiers are domain dependent [13], [14]. In praxis this leads to necessity of separate corpora for particular objects of evaluation, source types or languages. Hence, methods which try to increase the effectiveness of the annotation tools, like active learning, are becoming interesting.

The basic idea of active learning is the selection of unlabeled examples for manual annotation on the basis of their informativeness for the classification task [15]. The particular active learning methods differ in the way how this informativeness is calculated. A comprehensive overview of active learning methods in the context of natural language processing can be found in the work of Olson [15]. Despite the popularity of sentiment analysis in last years, the number of works related to the possibilities of their improvement by utilization of active learning is small. Boiy and Moens[16] in their work used three active learning strategies for sentiment analysis of blogs, reviews and forum contributions. The first of them was the classical uncertainty sampling based on the uncertainty of the classification of unlabeled data, which performed similarly as the random strategy. The second strategy was based on relevance sampling which uses examples most likely to be low-represented class members in order to increase the number of examples from class where there are too few examples. The third method was the *kernel farthest first* which uses examples farthest from the already labeled examples; however this strategy performed worse than the random sampling.

The problem of imbalanced samples is addressed by Li et al. [17], where an active learning approach for imbalanced sentiment classification tasks was proposed. They used complementary classifiers where the first one was used to get most certain examples from both classes and the second to get most uncertain examples from the minority class for manual annotation, while the classifiers were trained with two disjoint feature subspaces. Dasgupta and Ng[14] used active learning based on SVM where the classifier was trained on automatically labeled unambiguous reviews to identify ambiguous reviews for manual labeling and they confirmed the increase of annotation effectiveness. Zhou et al. [13] used active learning based on semi-supervised learning algorithm called active deep network. The above mentioned works computed the informativeness of examples using the classifiers uncertainty. However, there are some specifics of sentiment analysis which should be considered in the context of active learning. One of them is the possibility to use the above mentioned sentiment dictionaries, which represents an alternative to the informativeness computation methods based on the classifiers uncertainty. Therefore the aim of this work is to verify the possibilities of the use of methods from both mentioned groups of active learning methods considering their applicability in annotation tools.

3 Active Learning Strategies for Sentiment Analysis

On the basis of previous research we selected six active learning strategies applicable in the context of sentiment analysis, which were evaluated in a series of experiments. The first group of methods is represented by methods which use the confidence

estimates for particular classes based on results of classification models, which is a common method also for other application domains of active learning. This group is represented by the first three of the bellow mentioned strategies, where the informativeness was computed according to the equation 1. The equation is based on the information entropy using the posterior probability of the positive class $\hat{P}(C_a|X)$ and its supplement to one for the negative class for the feature vector X of the classified document.

$$Inf = -\hat{P}(C_a|X)\, log_2\, \hat{P}(C_a|X) - (1 - \hat{P}(C_a|X))\, log_2(1 - \hat{P}(C_a|X)) \quad (1)$$

Active learning strategy based on SVM. This active learning strategy uses the probability estimates of target classes based on SVM, which is well performing method for sentiment analysis and it is commonly used for active learning tasks. In this strategy the model trained on the corpus of annotated documents was used to classify the unlabeled documents. The informativeness of the unlabeled examples is then computed on the basis of difference between the probabilities of target classes. The probability of the positive class for the equation 1 is computed according to the equation 2, where $\hat{P}(C_a|X)$ is the posterior probability for the positive class for vector X and $f(X)$ is the output of the SVM. The probability of the negative class is computed as its supplement.

$$\hat{P}(C_a|X) = 1/(1 + exp(-f(X))) \quad (2)$$

The problem of this strategy is the relatively large amount of documents which don't contain any of the words from the model trained on corpus. In this case the computed probability equals for both classes; however these documents are not considered to be the best candidates for annotation. Hence these documents wouldn't be selected for annotation.

Active learning strategy based on Naïve Bayes classifier. Within the methods of machine learning NBC is another commonly used method for sentiment analysis. Also here is the informativeness of the document evaluated on the basis of equation 1. The probability of the positive class is computed on the basis of equation 3.

$$\hat{P}(C_a|X) = \hat{P}(X|C_a).\hat{P}(C_a)/\hat{P}(X) \quad (3)$$

Active learning strategy based on external model. The above mentioned strategies use classifiers trained on the corpora of annotated documents created during prior iterations. This strategy uses available corpora of annotated documents for sentiment analysis. An example of such a corpus, commonly used for evaluation of sentiment analysis methods, is the movie review corpus from the work of Pang and Lee [10], described in more details in the next section. The created model is then used to compute the classification uncertainty as in the previous strategies. The main issue in the utilization of these corpora is the mentioned domain dependency. On the other hand the whole pool of unlabeled documents is evaluated only once.

Active learning strategy based on external dictionaries. This and the following strategies are based on using of dictionaries of positive and negative words. An example sentiment dictionary is the commonly used MPQA Subjectivity Lexicon [18], which was used also in this work. From this dictionary we extracted a list of positive

and negative words. The number of occurrences of positive words N_p and negative words N_n in each document was used to compute the values of informativeness according to the equation 4. This strategy as well as the following strategies is based on assumption that documents containing words with opposite sentiment are more informative.

$$Inf = -\frac{N_p}{N_p+N_n} log_2 \frac{N_p}{N_p+N_n} - \frac{N_n}{N_p+N_n} log_2 \frac{N_n}{N_p+N_n} \quad (4)$$

Active learning strategy based on offline generated dictionaries. As it was mentioned in the related works, there are many methods which can be used to get dictionaries of positive and negative words, while the used method does this using the annotated documents. In our case information gain was used, which is a well performing method in feature extraction tasks. Features are in the case of text classification represented by vectors indicating the presence or absence of words or n-grams of words. Information gain was used to get the list of words relevant for sentiment analysis. The orientation of words was extracted as a simple sign of the difference between the number of occurrences of word in positive and negative evaluations. As a source of evaluations we used the corpus from the work of Pang and Lee [10]. Thus the sentiment dictionary was created before the active learning, hence the name offline.

Active learning strategy based on online generated dictionaries. This strategy is analogous with the previous one, while the difference is in the continuous (online) generation of dictionaries of positive and negative words on the basis of actual corpora of annotated evaluations in each iteration.

4 Experiments

The aim of the series of experiments was to evaluate the performance of the described active learning strategies in context of their application in annotation tools. For the comparison of evaluated strategies we used the experimental design depicted in Figure 1. At the beginning of the annotation we have a pool of unlabeled documents, from which we select examples for annotation. From this pool are in each iteration selected documents for annotation. The selection of documents is realized on the basis of compared active learning strategies, while in the first iteration the documents are selected randomly. Annotated documents are subsequently added to the corpus. As it will be described in the next chapter, the pool was in reality created by documents with known polarity, but the documents' polarity was not taken into account until adding them to the corpus. This solution was used to simulate the process of the real annotation. The number of documents added in each iteration to the corpus was 20 and in experiments were realized 100 iterations, i.e. the maximal size of corpora was 2000 documents. This corpora were used in each iteration to train the classifiers, which accuracy was evaluated on dedicated test set. The achieved accuracies correspond with the quality of the created corpus and the active learned strategy used to its creation. For the creation of classifiers we used SVM with linear kernel and C equal to 0. In parallel the informativeness of documents in pool was evaluated in case of strategies where this is done in each iteration. These values are then used to select

new documents for annotation in the next iteration. The whole process finishes by achieving the required size of corpus. Depending on the strategy sentiment dictionaries and external classifiers should be used. It should be also mentioned that we used binary vector representation of documents and each word was stemmed using the Porter stemmer. Besides the compared active learning strategies we used the baseline strategy, which is a simple random selection of documents from the pool of unlabeled documents.

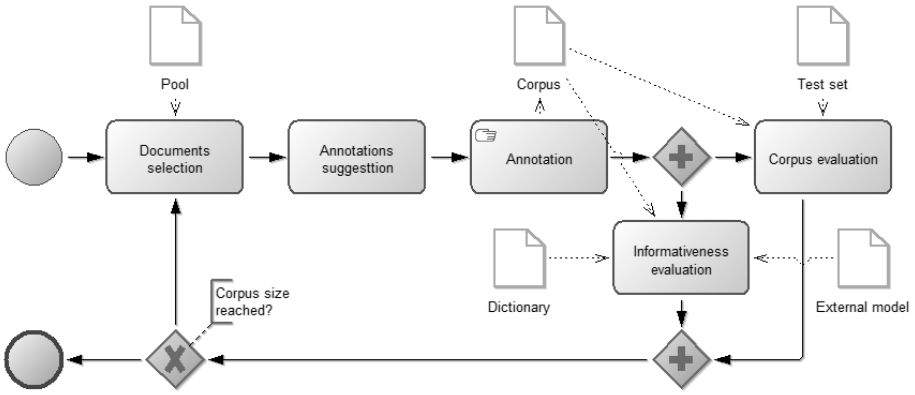


Fig. 1. Experimental design

4.1 Samples

For evaluation of possibilities of particular active learning strategies it was necessary to simulate the process of real annotation. Possible resources of evaluations with corresponding numerical evaluation representing the annotation are product review sites. For this purpose we crawled and parsed reviews from Reviewcenter[5]. From the original set of reviews a total of 17,240 were selected, with half made up by negative reviews (an evaluation from the interval 0 to 1) and half by positive reviews (evaluations equal to 5). The sample was created by random selection of reviews from different fields, with the aim of to create a balanced sample in respect to reviews orientation. From this sample was by stratified random sampling created a test set of evaluations with 1000 documents. The rest of the evaluations were used as a pool of unlabeled documents. The corpora used for training of classifiers were built on the fly by adding new documents in each iteration according to the informativeness of documents evaluated according to active learning strategies.

The data sample used for the active learning strategy based on external model and offline generated dictionaries was created from the corpus used by Pang and Lee [10]. It is a corpus made up of 1000 positive and 1000 negative reviews of films. This data sample is one of the most commonly used samples for sentiment analysis.

[5] http://www.reviewcentre.com/

4.2 Results

The results achieved by using of particular active learning strategies are depicted in graphs on figures 2 and 3. Axis x represents the sizes of corpora, whereas axis y represents the achieved classification accuracy on the test set. Besides the accuracy values achieved by compared strategies they are depicted also the accuracy values for the baseline strategy as well as the accuracy achieved by using the whole pool of documents for training, which equals 93,6%. The maximal accuracy achieved by the baseline strategy was 74% by using the maximal size of corpus.

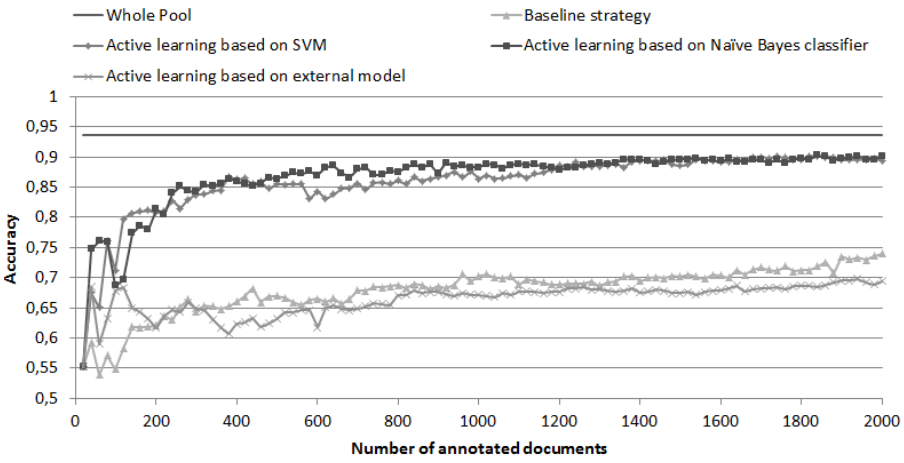


Fig. 2. Results for methods using classifier uncertainty

In Figure 2 are depicted the achieved accuracies for methods using the class probability estimates based on classifiers, i.e. active learning strategies based on SVM, NBC and external model. The best results were achieved by using active learning strategy based on SVM. The maximal achieved accuracy for this strategy was 90,2%. Similar accuracies were achieved also by using the active learning strategy based on NBC. The maximal achieved accuracy for this strategy was 90,3%. For both strategies were the values of accuracy significantly higher in comparison to baseline strategy. In case of active learning based on external model wasn't achieved increase of accuracy.

In Figure 3 are depicted the achieved accuracies for methods using sentiment dictionaries, i.e. active learning strategy based on external dictionaries and offline and online generated dictionaries. The best results were achieved by using active learning based on external dictionaries. The maximal achieved accuracy for this strategy was 84,7%. Active learning based on offline generated dictionaries achieved worse results. The maximal achieved accuracy for this strategy was 78,1%. Lowest accuracy was achieved by active learning strategy based on online generated dictionaries. Also in case of these methods were the achieved accuracies better than in case of baseline strategy, however the increase of accuracy for these methods was not so significant.It should be also mentioned that the accuracy increase for these strategies was more significant in first iterations and in some cases the accuracy even decreased after adding new documents.

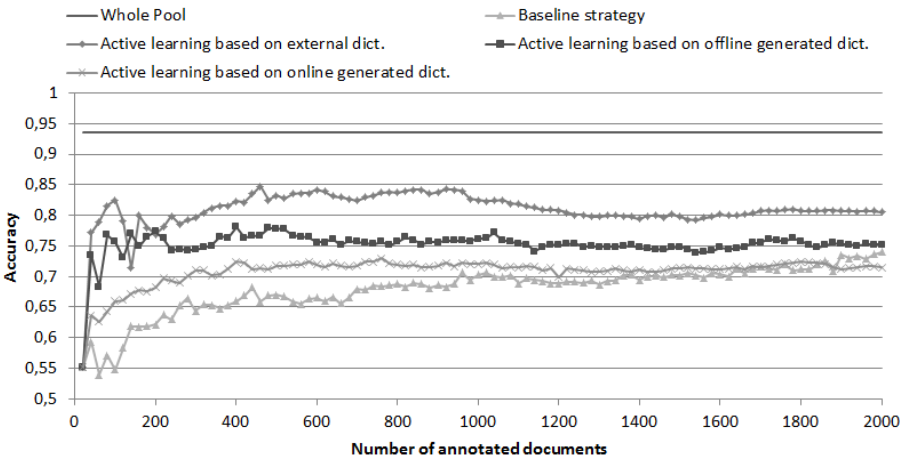


Fig. 3. Results for methods using sentiment dictionaries

5 Conclusions

The achieved results verified the efficiency of active learning methods in process of documents annotation for the needs of sentiment analysis. From the active learning strategies based on classifiers uncertainty active learning based on SVM has been shown as the best performing strategy. One of its advantages is the independency on quality of external resources. Its disadvantage is the necessity of model creation after each iteration. From the active learning strategies based on dictionaries the strategy using an external dictionary achieved the best performance. The advantage of this strategy is in the simplicity of its application. Moreover the dictionaries can be used for highlighting of words in annotated document and improve the efficiency of annotation. From the point of view of the application of these strategies in annotation tools the combination of both types of strategies should be useful, where in the initial phase of annotation the dictionary based methods can be used which will be then replaced by classifiers with uncertainty based methods. In our future work we will design such kind of combined annotation supporting active learning method and adjust it also for the purpose of aspect-based sentiment analysis.

Acknowledgment. This work was partially supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/1147/12 and partially by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

1. Koncz, P., Paralic, J.: An approach to feature selection for sentiment analysis. In: 15th IEEE International Conference on Intelligent Engineering Systems (INES 2011), pp. 357–362 (2011)
2. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool (2012)

3. Koncz, P., Paralič, J.: Automated creation of corpora for the needs of sentiment analysis. In: 3rd RapidMiner Community Meeting and Conference (RCOMM 2012), Aachen, pp. 107–113. Shaker Verlag, Aachen (2012)
4. Kamps, J., Marx, M., Mokken, R.J., de Rijke, M.: Using WordNet to Measure Semantic Orientations of Adjectives. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 1115–1118 (2004)
5. Kim, S.-M., Hovy, E.: Automatic Detection of Opinion Bearing Words and Sentences. In: Proceedings of the Second International Joint Conference on Natural Language Processing (JCNLP 2005), pp. 61–66 (2005)
6. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 79–86 (2002)
7. Abbasi, A., Chen, H., Salem, A.: Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Trans. Inf. Syst. 26(3), 1–34 (2008)
8. Abbasi, A., France, S., Zhang, Z., Chen, H.: Selecting Attributes for Sentiment Classification Using Feature Relation Networks. IEEE Transactions on Knowledge and Data Engineering 23(3), 447–462 (2011)
9. Prabowo, R., Thelwall, M.: Sentiment analysis: A combined approach. Journal of Informetrics 3(2), 143–157 (2009)
10. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 271. Association for Computational Linguistics, Barcelona (2004)
11. Xia, R., Zong, C., Li, S.: Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences 181(6), 1138–1152 (2011)
12. Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Winkler, S.M., Schaller, S., Holzinger, A.: On Text Preprocessing for Opinion Mining Outside of Laboratory Environments. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 618–629. Springer, Heidelberg (2012)
13. Zhou, S., Chen, Q., Wang, X.: Active deep learning method for semi-supervised sentiment classification. Neurocomputing (May 2013)
14. Dasgupta, S., Ng, V.: Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language, pp. 701–709 (2009)
15. Olsson, F.: A literature survey of active machine learning in the context of natural language processing SE - SICS Technical Report. Swedish Institute of Computer Science, Box 1263, SE-164 29 Kista, Sweden (2009)
16. Boiy, E., Moens, M.-F.: A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval 12(5), 526–558 (2008)
17. Li, S., Ju, S., Zhou, G., Li, X.: Active learning for imbalanced sentiment classification. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 139–148 (2012)
18. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2005), pp. 347–354 (2005)

On Graph Entropy Measures for Knowledge Discovery from Publication Network Data

Andreas Holzinger[1], Bernhard Ofner[1], Christof Stocker[1], André Calero Valdez[2], Anne Kathrin Schaar[2], Martina Ziefle[2], and Matthias Dehmer[3]

[1] Medical University Graz, A-8036 Graz, Austria
Institute for Medical Informatics, Statistics & Documentation, Research Unit Human-Computer Interaction
{a.holzinger,b.ofner,c.stocker}@hci4all.at
[2] Human-Computer Interaction Center, RWTH Aachen University, Germany
{calero-valdez,schaar,ziefle}@comm.rwth-aachen.de
[3] Institute for Bioinformatics and Translational Research, UMIT Tyrol, Austria
matthias.dehmer@umit.at

Abstract. Many research problems are extremely complex, making interdisciplinary knowledge a necessity; consequently cooperative work in mixed teams is a common and increasing research procedure. In this paper, we evaluated information-theoretic network measures on publication networks. For the experiments described in this paper we used the network of excellence from the RWTH Aachen University, described in [1]. Those measures can be understood as graph complexity measures, which evaluate the structural complexity based on the corresponding concept. We see that it is challenging to generalize such results towards different measures as every measure captures structural information differently and, hence, leads to a different entropy value. This calls for exploring the structural interpretation of a graph measure [2] which has been a challenging problem.

Keywords: Network Measures, Graph Entropy, structural information, graph complexity measures, structural complexity.

1 Introduction and Motivation for Research

Tradition in the history of science emphasizes the role of the individual genius in scientific discovery [3]. However, there is an ongoing trend away from such an individual based model of scientific advance towards a networked team model [4]. Teams can bring-in greater collective knowledge; however, the most convincing factor is, that multidisciplinary teams are able to maintain an integration and appraisal of different fields, which often provides an atmosphere to foster different perspectives and opinions; and this often stimulates novel ideas and enables a fresh look on methodologies to put these ideas into business [5].

This is mainly due to the fact that many research problems, e.g. in the life sciences, are highly complex, so that know-how from different disciplines is necessary. Cooperative work in cross-disciplinary teams is thus of increasing interest.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 354–362, 2013.

Consequently, mixed-node publication network graphs can be used to get insights into social structures of such research groups, but elucidating the elements of cooperation in a network graph reveals more than simple co-authorship graphs, especially as a performance metric of interdisciplinarity [1].

However, before we can select measures to improve communication effectiveness or interpersonal relationships, it is necessary to determine which factors contribute to the interdisciplinary success and furthermore what constitutes interdisciplinary success. Moreover, it is quite important to understand the measures in depth, i.e., what kind of structural information they detect [6], [7], [8], [9].

2 Methods and Materials

As in previous work [1] we use a mixed node graph in order to analyze publication behavior. We create a reduced mixed node public network to demonstrate the research efforts of an interdisciplinary research cluster at the RWTH Aachen University. Typically bibliometric data is visualized using co-authorship graphs, leaving out the element of the interaction (i.e. the publication). The use of mixed node publication network graphs allows a graph to contain more information (than a co-authorship graph) and can easily be reduced to one by using an injective mapping function. This type of graph allows fast human analysis of interdisciplinarity by explicating the authors tension between his discipline and his (possibly interdisciplinary) publications. When visualized properly this graph will match the users mental model, which is important in recognition tasks [10]. In our particular case we use the reduced Graph G_r.

2.1 Construction of the Network Graph

The network graph G_r is constructed equally as in previous work [1] with two node types. A node in this case represents either an author (A-Node) or a publication (P-Node). Nonetheless both node types (i.e. vertices) are not regarded as differently from a graph theory point of view. We define the two sets representing authors and publications as follows:

$$A = \{a \mid \text{a is author in cluster of excellence at RWTH}\} \tag{1}$$

$$P = \{p \mid p \text{ is a publication funded by the cluster written by any } a \in A\} \tag{2}$$

We also define two vertex-mappings f_a and $f_p$and two sets of vertices V_1 and V_2 as follows:

$$f_a : A \to V_1, f_a(a) = v; a \in A \wedge v \in V_1 \tag{3}$$

$$f_p : P \to V_2, f_p(p) = v; p \in P \wedge v \in V_2 \tag{4}$$

$$\text{with } V_1 \cap V_2 = \varnothing \tag{5}$$

These function represent mappings of authors and publications to vertices, and if inverted finding the "meaning" of a vertice. We define the sets E as all edges between authors and publications, when an author has written a publication:

$$E = \{e \mid e = (v_1, v_2), v_1 \in V_1 \wedge v_2 \in V_2 \wedge f_a^{-1}(v_1) \text{ is author of } f_p^{-1}(v_2)\} \quad (6)$$

We use the graph definition for Gr for our analyses:

$$G_r = \{(V, E) \mid V = V_1 \cup V_2\} \quad (7)$$

This bipartite graph can be visualized using standard graph visualization tools. In order to enable analysis by a human person the graph needs be lain out visually. Graph visualization is done with both Gephi [11] and D3JS (www.d3js.org). In this case 2D-spatial mapping is performed by using force-based algorithms.

For our visualization we set the size of nodes according to their corresponding degree and applied a grayscale color scheme based on betweenness centrality (see Fig. 1). Fig. 2 shows the two different node sets. White nodes denote authors, gray nodes denote publications. Both graphs share the same layout (Force Atlas 2, linlog mode, no overlap, scaling=0.3, gravity=1) and node sizes (min=10, max=50). Nodes sizes were chosen as node degrees. Authors with more publications are bigger, as well as publications with more authors are bigger.

Using force-based algorithms has the following consequences for graph visualization:

- All nodes are attracted to the center, but repel each other.
- All nodes that are connected by an edge attract each other (i.e. an author and his publication).

This allows the following visual conclusions:

- Two A-Nodes are spatially closer if they publish together (triangle inequality).

2.2 On Graph Entropies

The open source graph visualization tool Gephi allows for several different graph analyses of network graphs. Traditionally these are used with social network graphs (i.e. co-authorship graphs). Interpretation of graph statistics must be reevaluated for mixed node graphs. Graph statistics that are of interest in regard to publication networks are:

- Network Entropies have been developed to determine the structural information content of a graph [7], [2]. We have to mention that the term network entropy cannot be uniquely defined. A reason for this is that by using Shannon's entropy [12], [13], [14] the probability distribution cannot be assigned to a graph uniquely. In the scientific literature, two major classes have been reported [7], [6], [15]:

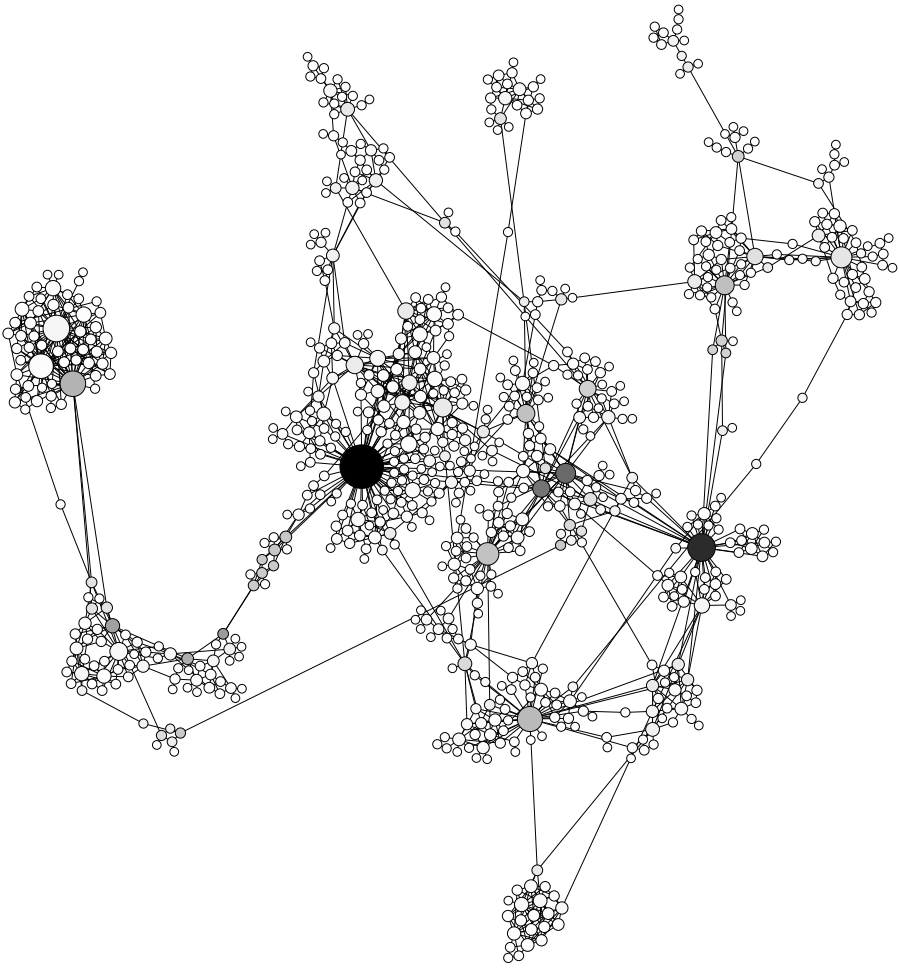


Fig. 1. G_r of publication data of the excellence network from the RWTH Aachen. The node size shows the node degree whereas the node color shows the betweenness centrality. Darker color means higher centrality.

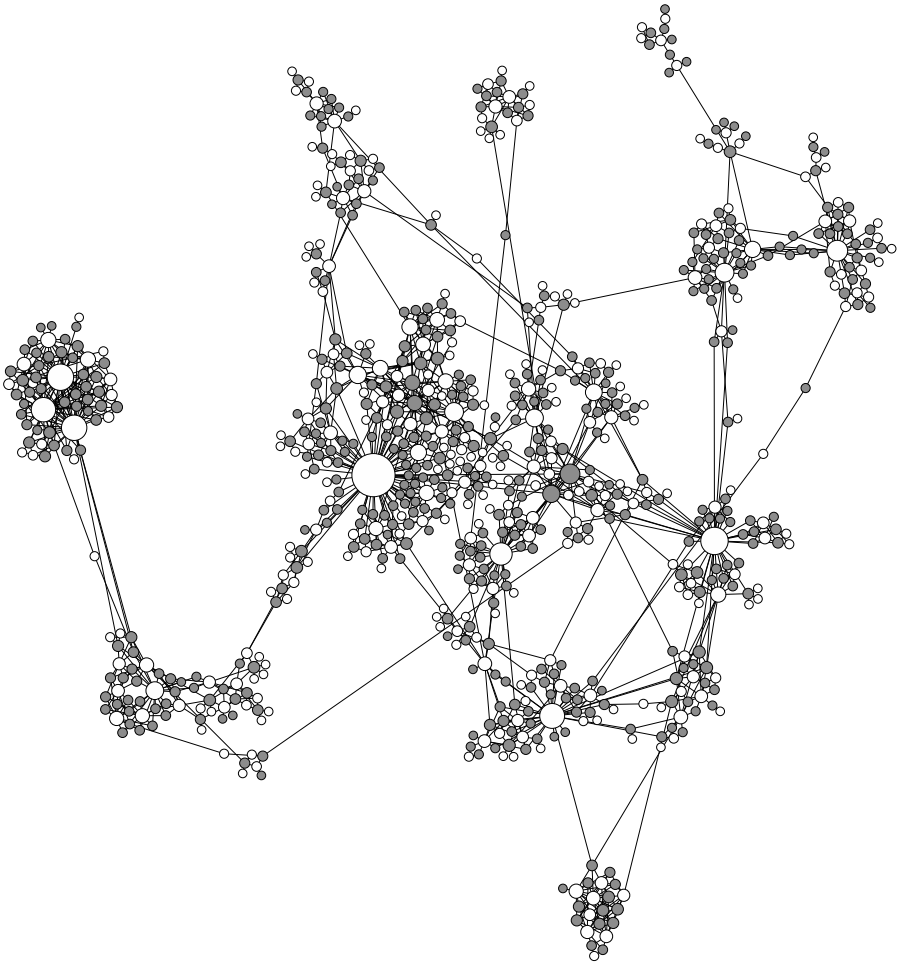


Fig. 2. G_r of publication data of the excellence network from the RWTH Aachen. The node size shows the betweenness centrality. White nodes denote authors, gray nodes denote publications.

1. Information-theoretic measures for graphs which are based on a graph invariant X (e.g., vertex degrees, distances etc.) and an equivalence criterion [13]. By starting from an arbitrary graph invariant X of a given graph and an equivalence criterion, we derive a partitioning. Thus, one can further derive a probability distribution. An example thereof is to partition the vertex degrees (abbreviated as $\delta(v)$) of a graph into equivalence classes, i.e., those classes only contain vertices with degree $i = 1, 2, ..., \max \delta(v)$, see e.g. [8].
2. Instead of determining partitions of elements based on a given invariant, Dehmer [6] developed an approach which is based on using so called information functionals. An information functional f is a mapping which maps sets of vertices to the positive reals. The main difference to partition-based measures (see previous item) is that we assign probability values to every individual vertex of a graph (and not to a partition), i.e.,

$$p^f(v_i) := \frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)} \tag{8}$$

As the probability values depend on the functional f, we infer a family of graph entropy measures

$$I_f(G) := -\sum_{i=1}^{|V|} p^f(v_i) \log p^f(v_i) \tag{9}$$

$|V|$ is the size of the vertex set of G. Those measures have been extensively discussed in [8].

Evidently, those information-theoretic graph measures can be interpreted as graph complexity measures.

The Graph G_r contains 796 nodes split into 323 authors and 473 publications linked by 1677 edges. Applying the Gephi graph analysis reveals the following statistics. The graph shows an average degree of 4.214 and a network diameter of 23. The average path length is 7.805 and graph density is .005. The graph only contains a single connected component.

3 Evaluation of the Graph by Using Network Entropies

In this section, we evaluate some information-theoretic network measures (graph entropies) on the given Excellence Network. To start, we briefly characterize this network by stating some graph-theoretical measures.

We evaluated the following graph entropies:

– A partition-based graph entropy measure called *topological information content* based on vertex orbits due to [13].

– Parametric graph entropies based on a special information functional f due to Dehmer [6]. The information functional we used is

$$f(v_i) := \sum_{k=1}^{\rho(G)} c_k |S_k(v_i, G)|, \text{ with } c_k > 0 \tag{10}$$

summing the product of both the size of the k-sphere (i.e. the amount of nodes in G with a distance of k from v_i given as $|S_k(v_i, G)|$) and arbitrary positive correction coefficients c_k for all possible k from 1 to the diameter of the graph G. The resulting graph entropies have been defined by

$$I_f := -\sum_{i=1}^{|V|} p^f(v_i) \log p^f(v_i) \tag{11}$$

– Network entropy due to [16].
– Graph entropy based on the ER model for modeling random graphs [17].

Table 1. Calculated graph entropies

Method	**Symbol**	**Graph Entropy**
Topological information content [13]	I_{mowsh}	9.031485
Parametric graph entropies [6]	I_{dehm}	9.6258
Network entropy due to [16]	I_{valv}	0.3095548
Graph entropy based on the ER model [17]	I_{wang}	15090.71

We can note that the used graph entropies evaluate the complexity of our network differently. Here we will explore this problem with in illustrative exmaple, namely by considering the measures $I_{\text{mowsh}} < I_{\text{dehm}}$. In this context, the inequality $I_{\text{mowsh}} < I_{\text{dehm}}$ can be understood by the fact those entropies have been defined on different concepts.

As mentioned, I_{mowsh} is based upon the automorphism group of a graph and, therefore, can be interpreted as a symmetry measure. This measure vanishes if all vertices are located in only one orbit. By contrast, the measure is maximal ($= \log_2(|V|)$) if the input graph equals the so-called identity graph; that means all vertex orbits are singleton sets. In our case, we obtain $I_{\text{mowsh}} = 9.0315 < \log_2(796) = 9.6366$ and conclude that according to the definition of I_{mowsh}, the excellence network is rather unsymmetrical.

Instead, the entropy I_{dehm} characterizes the diversity of the vertices in terms of their neighborhood, see [7]. The higher the value of I_{dehm}, the less topologically different vertices are in the graph and, finally, the higher is the inner symmetry of our excellence network. Again, maximum entropy for our network equals $\log_2(796) = 9.6366$. Based on the fact that for the complete graph K, $I_{\text{dehm}}(K_n) = \log(n)$ holds, we conclude from the result $I_{\text{dehm}} = 9.6258$ that the excellence network is highly symmetrical and connected and could theoretically be obtained by deleting edges from K_{796}.

The interpretation of the results for I_{valv} and I_{wang} can be done similarly by arguing based on their definitions.

4 Discussion

Different entropy measure deliver different results because they are based on different graph properties. When using the aforementioned entropy measures in a mixed-node publication graph measures of symmetry I_{dehm} (based on vertex neighborhood diversity) or I_{mowsh} (based on the graph automorphism) deliver different measures of entropy. Interpreted we could say, that authors/publications are similar in regard to their neighborhoods (i.e. authors show similar publication behavior, publications show similar author structures) but the whole graph shows low measures of automorphism-based symmetry to itself. This could mean authors or publications can not be exchanged for one another without changing basic properties of the graph. But since authors and publications are used in the same vertex set there are also implications of interpretation between these sets. For example a graph isomorphisms that maps vertices from V_1 to V_2 should not be included in the measure, because they are not intelligible from an interpretation point of view. New measures of entropy specialized for mixed-node graphs are required to accurately measure graph properties in such graphs.

5 Conclusion

In this paper, we evaluated information-theoretic network measures on publication networks. In our case, we used the excellence network from the RWTH Aachen, described in [1]. Those measures can be understood as graph complexity measures which evaluate the structural complexity based on the corresponding concept.

A possible useful interpretation of these measures could be applied in understanding differences in subgraphs of a cluster. For example one could apply community detection algorithms and compare entropy measures of such detected communities. Relating these data to social measures (e.g. balanced score card data) of sub-communities could be used as indicators of collaboration success or lack thereof, as proposed in [18] and [19].

Nonetheless we see that it is challenging to generalize such results towards different measures as every measure captures structural information differently and, hence, leads to a different entropy value. This calls for exploring the *structural interpretation* of a graph measure [2] which has been a challenging problem.

Acknowledgments. We cordially thank the Aachen House of Production for allowing us to analyze their publication data. This research was partially funded by the Excellence Initiative of the German federal and state governments.

References

1. Calero Valdez, A., Schaar, A.K., Ziefle, M., Holzinger, A., Jeschke, S., Brecher, C.: Using mixed node publication network graphs for analyzing success in interdisciplinary teams. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 606–617. Springer, Heidelberg (2012)

2. Dehmer, M.: Information theory of networks. Symmetry 3(4), 767–779 (2011)
3. Merton, R.: The Matthew Effect in Science: The reward and communication systems of science are considered. Science 159(3810), 56–63 (1968)
4. Wuchty, S., Jones, B., Uzzi, B.: The increasing dominance of teams in production of knowledge. Science 316(5827), 1036–1039 (2007)
5. Holzinger, A.: Successful Management of Research Development. BoD–Books on Demand (2011)
6. Dehmer, M.: Information processing in complex networks: Graph entropy and information functionals. Appl. Math. Comput. 201(1-2), 82–94 (2008)
7. Dehmer, M., Varmuza, K., Borgert, S., Emmert-Streib, F.: On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures. Journal of Chemical Information and Modeling 49, 1655–1663 (2009)
8. Dehmer, M., Mowshowitz, A.: A history of graph entropy measures. Inf. Sci. 181(1), 57–78 (2011)
9. Holzinger, A., Stocker, C., Bruschi, M., Auinger, A., Silva, H., Gamboa, H., Fred, A.: On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 646–657. Springer, Heidelberg (2012)
10. Calero Valdez, A., Ziefle, M., Alagöz, F., Holzinger, A.: Mental models of menu structures in diabetes assistants. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010, Part II. LNCS, vol. 6180, pp. 584–591. Springer, Heidelberg (2010)
11. Salotti, J., Plantevit, M., Robardet, C., Boulicaut, J.F.: Supporting the Discovery of Relevant Topological Patterns in Attributed Graphs (December 2012), Demo Session of the IEEE International Conference on Data Mining (IEEE ICDM 2012)
12. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal 27 (1948)
13. Mowshowitz, A.: Entropy and the complexity of graphs: I. An index of the relative complexity of a graph 30, 175–204 (1968)
14. Holzinger, A., Stocker, C., Peischl, B., Simonic, K.M.: On using entropy for enhancing handwriting preprocessing. Entropy 14(11), 2324–2350 (2012)
15. Mowshowitz, A., Dehmer, M.: Entropy and the complexity of graphs revisited. Entropy 14(3), 559–570 (2012)
16. Solé, R., Valverde, S.: Information Theory of Complex Networks: On Evolution and Architectural Constraints. In: Ben-Naim, E., Frauenfelder, H., Toroczkai, Z. (eds.) Complex Networks. Lecture Notes in Physics, vol. 650, pp. 189–207. Springer, Heidelberg (2004)
17. Ji, L., Bing-Hong, W., Wen-Xu, W., Tao, Z.: Network entropy based on topology configuration and its computation to random networks. Chinese Physics Letters 25(11), 4177 (2008)
18. Jooss, C., Welter, F., Leisten, I., Richert, A., Schaar, A., Calero Valdez, A., Nick, E., Prahl, U., Jansen, U., Schulz, W., et al.: Scientific cooperation engineering in the cluster of excellence integrative production technology for high-wage countries at rwth aachen university. In: ICERI 2012 Proceedings, pp. 3842–3846 (2012)
19. Schaar, A.K., Calero Valdez, A., Ziefle, M.: Publication network visualisation as an approach for interdisciplinary innovation management. In: IEEE Professional Communication Conference (IPCC) (2013)

Visualization Support for Multi-criteria Decision Making in Geographic Information Retrieval

Chandan Kumar[1], Wilko Heuten[2], and Susanne Boll[1]

[1] University of Oldenburg, Oldenburg, Germany
{chandan.kumar,susanne.boll}@uni-oldenburg.de
[2] OFFIS - Institute for Information Technology, Oldenburg, Germany
wilko.heuten@offis.de

Abstract. The goal of geographic information retrieval (GIR) is to provide information about geo-entities to end-users and assist their spatial decision making. In the current means of GIR interfaces, users could easily visualize the geo-entities of interest on a map interface via sequential querying or browsing of individual categories. However, there are several decision making scenarios when the user needs to explore and investigate the geospatial database with multiple criteria of interests, which is not well supported by the sequential querying or browsing functionality of current GIR interfaces. There is a need for more sophisticated visual interfaces to enable end-users in discovering knowledge hidden in multi-dimensional geospatial databases. In this paper we discuss some of the HCI issues in realizing such multi-criteria decision making scenario based on the user requirement analysis. To tackle the human centered aspects we propose different heatmap based interfaces to support multi-criteria visualizations in GIR, i.e., to facilitate the knowledge based exploration of geospatial databases with less information overload.

Keywords: Local Search, Geographic Information Retrieval, Geovisualization, User Interfaces, User-Centered Design, Heatmaps, Grids, Voronoi diagram.

1 Introduction

Geo-related information is one of the basic needs of citizens to understand the local infrastructure and to satisfy their spatial information requirements [1]. Spatial databases of georeferenced entities, documents and Web pages are a huge information source for geographic locations and entities. A GIR system supports end-users to search into these spatial databases [11], its goal is to process the information request of end-users and provide the results to satisfy their spatial information need. The current GIR interfaces are able to serve comparably simple requests, e.g., a search for "*Restaurant in New York*" in Google Maps[1] will return all restaurants in the downtown New York. However,

[1] http://www.maps.google.com

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 363–375, 2013.

in the current information age, the spatial information need of an end-user is often much more complex than such simple, aforementioned queries.

In several scenarios of spatial decision making users look for multiple criteria of interest simultaneously, which is not very convenient with the sequential querying and categorization approach of current GIR interfaces, e.g., somebody who has to move to a new city and is looking for a new living place. In this example the user will most likely come up with a variety of different criteria of geo-entities that he/she would like to visualize it together, e.g., availability of shopping facilities, medical facilities, and a good connection to public transport. Existing GIR systems fail to satisfy such complex information needs of end-users, even though the associated geospatial databases contain all these multi-dimensional information. The need is to support geographic retrieval with appropriate visualization methods that could assist users to explore the geospatial databases more effectively to discover the hidden spatial knowledge.

In this paper we argue the need of visualization methods for GIR to assist the complex decision making task of end-users. We conducted a user study to understand the requirements of end-users for GIR decision support systems, where we presented some spatial decision making scenarios such as moving to a new region or city. The insights of our study motivated us to design different interfaces which can support users in visualizing multiple criteria on the map with less information overload. We primarily focus on the heatmap visualizations (which have been simple yet effective phenomena in visualization research [16,6] for the information representation with less overload), and proposed various GIR interfaces using aggregated computations of geo-entities with grid-based and voronoi divisions. The proposed interfaces operate with the effective visualizations to enhance the capability of geospatial database in providing useful knowledge about spatial regions, and assist end-users decision making task.

The rest of paper is organized as follows: First, we give an overview of the multi-criteria local search problem and discuss the related applications and research approaches in Section 2. Then in Section 3 we describe our user study which was conducted to characterize multi-criteria search over various scenarios of exploring urban regions. We propose our heatmap based visual interface for multi-criteria local search in Section 4. In this section we describe how grid-based and voronoi- divisions were employed to present the aggregated spatial information. Finally, Section 5 concludes the paper by discussing the contribution of the paper and directions for future work.

2 Background and Related Work

Geographic information retrieval or local search provides the information about geo-entities to end-users [1,11]. These geo-entities usually belong to a criteria/category or set of categories (shopping, education, sport, etc.). Several location-based services and yellowpages follow the categorical structure for the overview of geo-entities. In the current geo applications, users could view the geo-entities which belong to a particular category on the map, but realization of

multiple categories is usually not supported. So end-users are expected to accomplish decision making task through sequential querying/browsing of categories which could be extremely complex and time consuming task. There are a few commercial applications[2,3] which provide the end-user interface of few selected categories for a very specific task like hotel or apartment search. In comparison we focus on the generic local search for end-users, and to support their decision making through visualization of multiple categories on a map interface.

In the research community, multi-criteria spatial problems have been approached from the computation perspective [18,22] which is focused on distance and density estimation of locations with respect to different criteria. In comparison, we focus on visualization methods, i.e., how end-users perceive and analyze multiple criteria on local search interface for spatial decision making. Some geovisualization approaches [5,7] explore the problem of multi-criteria analysis, but they usually target specific domains like healthcare and support the decision analysis of a focused group of experts. Rinner and Heppleston [21] proposed geospatial multi-criteria evaluation for home buyers where decision criteria were based on: location, proximity, and direction. Even though their task has been guided by a similar scenario of spatial decision making like discussed in our work, the study was conducted for a focused group of real-estate agents and the contribution was more on the computation issues rather than the HCI and visualization aspects. In general geospatial decision making has been one of the main challenges and application of visual analytics [10]. For geo-related content visual analytics is prominently being used. There have been many visual analytical models and tools developed to support critical business decision making processes, but assisting lay users in their decision making is still a major research challenge.

In this paper we focus on the analysis of GIR visual interfaces with end-users. The research problem of user-centered analysis have been well investigated in visualization and human-computer interaction research[19,12,17], also in the spatial context of geovisualization[13,20]. Even though we follow the general guidelines from the aforementioned work, we particularly focus on the usability and end-user satisfaction of a heatmap based visualizations. It is common to use heatmaps, grid structure, and voronoi divisions, in many visualization approaches, but their impacts on the GIR interfaces to support user's multi-criteria decision making have not been explored before.

In this work we particularly augment the interfaces of our existing GIR system whose goal is to provide the access of geo-entities through spatial Web pages. Figure 1 shows the end-user interface of system which is very similar to the current local search services, i.e., markers on the map to show the geo-entities, and an associated ordered list of results. The associated geospatial database consist a huge collection of georeferenced Web pages through a focused Web crawler [3]. Identification and verification of the addresses are done against the

[2] `http://www.hipmunk.com/hotels/`
[3] `http://www.walksore.com/apartments/`

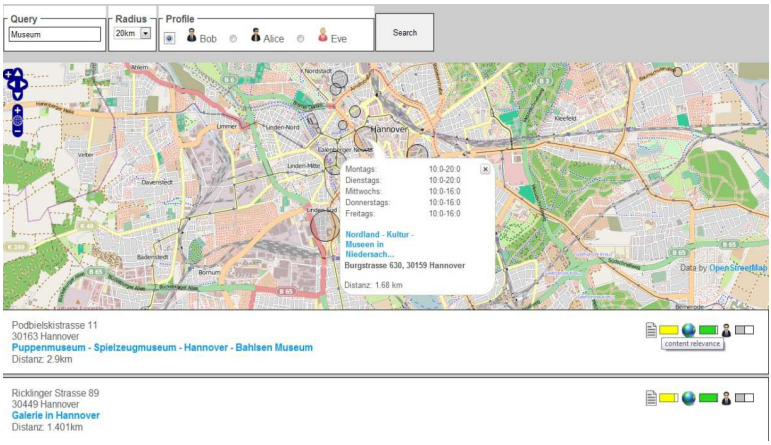


Fig. 1. A screenshot of our early GIR interface

street name database, which we fetched from OpenStreetMap[4] for the major cities of Germany. The system is primarily focused on relevance algorithms and computational ranking solutions [11,2] to process user requests and listing result documents in a ranked order.

3 User Requirement Analysis in Complex Spatial Decision Scenarios

The role of end-user is of prime importance in the successful design of Web interfaces, so the usability and HCI aspects need to be considered [9,17]. We follow the generic guidelines of user-centered design process to understand the context of use and to gather requirements. We conducted a semi-structured interview, where we asked people about experiences with existing systems, gathered insights, desires and ideas for improvements regarding GIR decision making tasks. Ten users participated in our study. The participants were aged between 33 (mean) and 5.6 (standard deviation). We asked the participants to think about how it feels to move to a different city. We thereby hoped that the participants can provide us with more thoughtful, inspiring information. Initially, we asked each participant to name the important criteria and aspects if they would like to move to a different city, or to open up a business in a city, or to compare different cities. Then we asked each participant on how an ideal user interface for these scenarios could look like. Thereby, also non-stationary devices could be considered. Participants were allowed to draw sketches with pen and paper. We then asked them to reflect on how they use existing tools, like Google Maps, the GIR interface (Figure 1), for these scenarios and if they fulfill their needs.

[4] `http://www.openstreetmap.org/`

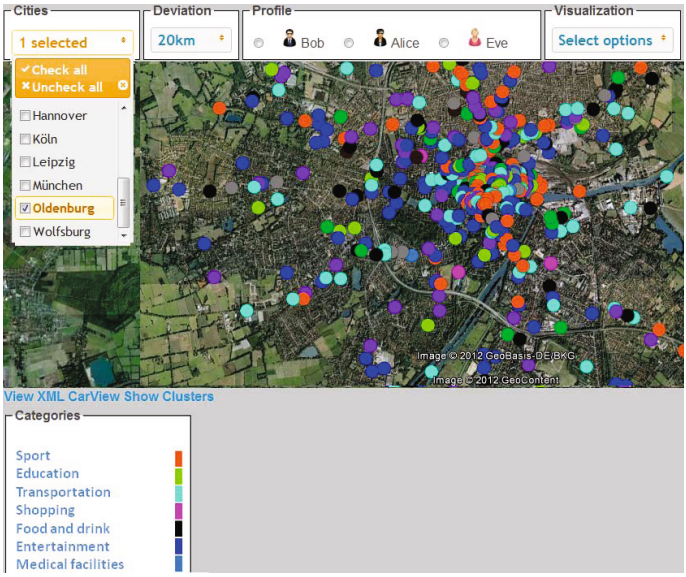


Fig. 2. Marker-based interface for multi-criteria exploration

Ultimately, we confronted the participants with an early functional prototype of our GIR interface for multi-criteria search (Figure 2) and asked them to share their thoughts by thinking aloud. The prototype is the easy adaption of the existing GIR interface to overcome the limitation of sequential querying and browsing. It supports multi-criteria selection through the colored association of markers on map. Each marker (circle) represents a geo-entity and follows the color scheme of its respected category.

We obtained the general perception that the user's requirements could be very diverse and dynamic. All study participants had their very own criteria in mind when discussing about the scenario, some criteria were contradictory, e.g., one participant wants to live as close as possible to work, while another participant wants to live as far away from work as possible to have a clear separation. Most of the participants stated that existing tools and services are not sufficient to fulfill these tasks in a simple and easy way. The current means of interfaces support only the sequential querying and searching of geo-entites, so the exploration with respect to multiple criteria of interests becomes an uneasy task. Few participants mentioned that they could achieve the task via several sequential queries. Moreover, they supported our point of view that this decision making process is very complex and the way to merge information in mind is demanding. Simplification in the information presentation, categorical and aggregated overview, easy and adaptive interaction were the main wishes of the users, we discuss these observations more specifically in the following.

Simplification of Information Presentation. Most of the users wished for easy and simple visualizations while being asked about their need in realizing an ideal geospatial decision support system. Participants were concerned from the overload of information from existing interfaces and the complexity of representation. The geospatial databases contain multivariate and complex attributes so the need is to summarize this huge information via easy visualizations. In order to support simple visualization, advance interaction methods could be useful to provide exploration capabilities to end user, i.e., to reduce the information loss.

Visualization at Category Level Is Important. We noticed that users tend to visualize the geo attribute at low granularity. While being asked to list the important geo attributes for relocation, users mention the geo-relevance parameters with a categorical overview, e.g. shopping, sports, education etc. The usual geo-data available to search systems are at the high granularity with real physical addresses of local entities. So the need is to represent the geovisualization of spatial database with low level category based details to simplify the initial visualization for end users.

Aggregation of Information and Knowledge-Based Support. In the scenarios such as geovisualization, integration of geo-spatial knowledge, learning of geo-spatially related entities in the search and visualization is required. There is a greater need of visual analysis tools to represent and reason with human knowledge and formalize the intersection between interaction and knowledge construction. From the user study we got the impression that the user is very willing to support these knowledge-based interfaces. In many of the drawn sketches we identified principles to assist such interfaces. Most prominently, the users wanted to set their general priorities and interests on some scales. These could serve as a starting point for building up more detailed knowledge-based interfaces. As open challenge we see how the user's set preferences and intelligent interface support work together, in particular since the user's interests might change over time or for each session.

Adaptive Interaction Techniques. Interfaces of spatial data usually demands interaction patterns like zooming, panning, rotating or selecting. The challenge is to provide users the scope to utilize these techniques to navigate the map and, thus, the geographic search results. The user should be able to mark areas, move to other areas, and manipulates the retrieved attributes till he/she reaches a deep understanding of the data and potentially gathers new knowledge. From the user study we got strong support that interaction techniques beyond traditional WIMP (windows, icons, menus, pointer) interfaces should be considered. We learned that the context is often completely different between users. It might be even different in various situations of the same user. E.g., the user's individual interaction flow might be different to the flow when the interests of a whole family are considered. These different profiles are very hard to infer from a

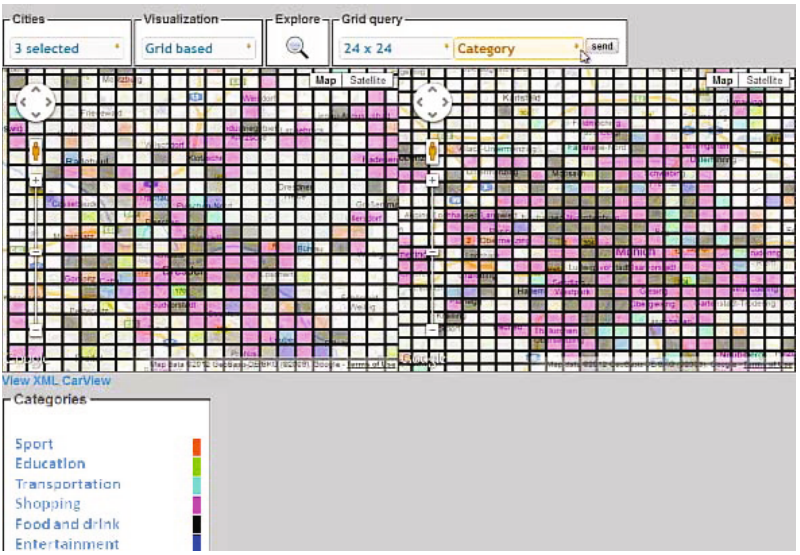


Fig. 3. Grid-based heatmap visualization for spatial exploration

technical point of view. However, the participants appreciated the different user profiles of the shown prototype (Figure 2). Thus, this is a significant challenge that needs to be addressed.

4 Visualizations to Support Multi-criteria Decision Making in GIR

Based on the gathered insights we are able to make some fundamental design decisions. We know that people will approach a spatial decision making task, like in the relocation scenario, with a map-based tool. Consequently, we decide that the main part of the user interface should be a map. This map should exactly behave like traditional map-based systems to match the user expectations regarding common interactions like zooming or panning. The user interface should use a simple language, should be easy to understand, should be clearly structured and further should avoid information overload. A trade-off was deemed optimal for the criteria selection process. The users should be allowed to adapt the criteria to their very individual needs. The user preferred criteria gathered in the requirement analysis guided our formulation of the list of category.

Our primary goal is to provide an appropriate visualization of the spatial data for GIR users. In visualization research, there have been many advance multivariate, high dimensional visualizations for analysts and decision makers which could be hard to interpret by lay users. Here we imply the heatmap based information presentation phenomena in the GIR interface, which is easily understandable and less prone to information overload. Heatmap visualization has been applied in various domains to represent spatial distributions and relevance [6].

Fig. 4. Ranked grid-based visualization

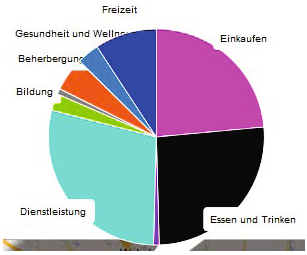


Fig. 5. Pie charts for user interaction

4.1 Grid-Based Heatmap Visualization for Spatial Exploration

For the aggregational view of geo-entities we discretize the map area using a grid raster [16], since we didn't intend to restrict user exploration to be constrained by administrative boundaries. We adapted the heatmap visualization for the distribution of categories across user-defined grid-based division of maps. Each grid cell conveys the categorical information. The color of each grid cell represents the category of largest allocation of geo-entities in the cell area. Figure 3 shows the grid-based visualization for two user selected cities with respect to 6 selected categories of interest. We can perceive that the visualization gives a

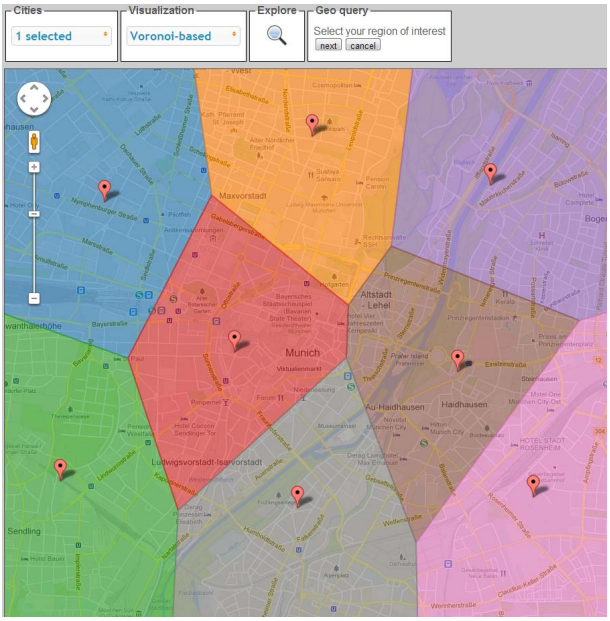


Fig. 6. Voronoi-based heatmap visualization

good overview and assessment on the distribution of geo located facilities across the region and cities, e.g., to identify the regions with better shopping facilities in the city.

Relevance Based Aggregation. Sometimes users need to rank the regions with respect to their criteria of interest, e.g., to find the best place to live or to open a business. We therefore employed a basic multi-criteria aggregation [21] to provide a spatially-related aggregated view on entity distribution with regard to multiple categories. Figure 4 shows the aggregated ranked version of the grid-based visualization. Here each grid cell represents the overall relevance with respect to user selected categories. The relevance score is computed based on the quantity and balanced distribution of geo-entities which belongs to the criteria of interests. The relevance of each grid cell G is based upon user selected category allocation $(C_1, C_2, .., C_n)$, where each category C_i has an importance based on the geo-entities which belong to category i. The relevance of a particular cell G is computed with the combined distribution of categories selected by user. We used a color scheme of six different green tones which differed in their transparency. While light colors represented low relevance, dark colors were used in order to indicate high relevance. The color scheme was chosen with the support of ColorBrewer[5]; an online color scheme recommendation tool based on the work of Brewer and Harrower [8].

[5] ColorBrewer: `http://www.colorbrewer.org`

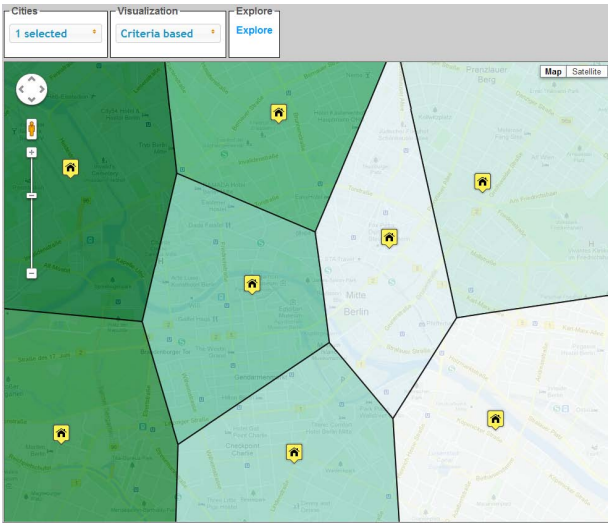


Fig. 7. Aggregated voronoi-based visualization

For the end-user interaction with grid and categories, we show the relevance of a grid with respect to selected categories using a pie chart (Figure 5). The pie charts provide the detailed information on categorical distribution of grid cells when the users click on a particular grid cell with mouse cursor.

4.2 Voronoi-Based Heatmap Visualization for User-Constrained Spatial Exploration

The aforementioned grid-based visualization divides the spatial regions in automated fix sized spaces, i.e. the end-user does not have the possibility of specifying personal location preferences as input. The user study also indicated the need of users in preselecting the desired locations and then compares them with respect to their criteria of interest. Figure 6 shows our transformation of heatmap visualization to a voronoi-based division to fulfill such requirement. Voronoi diagram [4] is a way of dividing space into a number of regions. A set of points (seeds) is specified beforehand and for each seed there will be a corresponding region consisting of all points closer to that seed than to any other. The regions are called voronoi cells. The color of each voronoi cell represents the category of largest allocation of geo-entities in the cell area.

Similar to the aggregated view of relevance in grid-based visualization, Figure 7 shows the ranked voronoi-based divisions. Here the relevance of a particular voronoi cell is computed not only based on the distribution of geo-entities; it's additionally the accumulation of root distance of the geo-entities from the selected home location. Figure 8 shows the distribution of geo-entities across the selected locations of city Leipzig with respect to categories: shopping, education and medical facilities. We could see that the most relevant region (voronoi cell

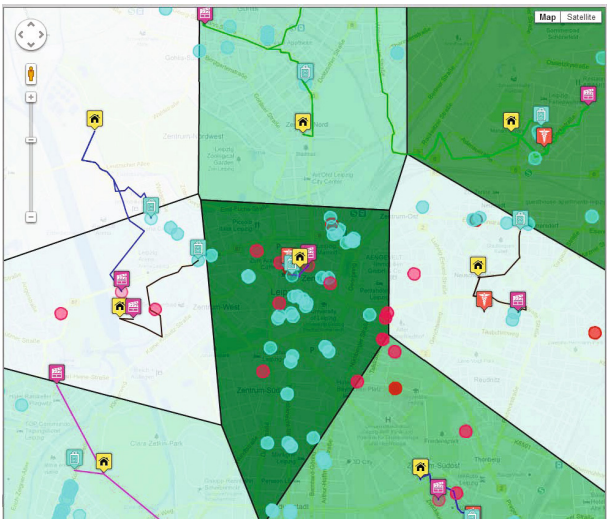


Fig. 8. Aggregated voronoi-based visualization with distribution of geo-entities

with the darkest green shade) has the better distribution of geo-entities and the corresponding home location is closely connected to the geo-entities of interest.

5 Conclusion and Future Work

Exploration of multiple criteria for location-based decision making is a common scenario of our daily life, but the current means of local search interface does not offer adequate support for such task. This was also evident from our study when the participant reported on the incapability of Google maps and given GIR interface to accomplish the task of exploring city with multiple criteria. In this paper we argue the need of interactive visual interfaces for a geographic information retrieval system, so the end-user can easily explore and interact with the spatial database. We presented an extension of current interfaces (with multiple placing of markers) to perform multi-criteria search. But the huge amount of markers presented in the marker condition carry the information overload and put high mental demand to perform the task of exploration and comparison of spatial regions. We proposed the use of heatmap visualization through grid and voronoi-based conditions to facilitate the knowledge based exploration of geospatial databases with less information overload. The proposed visualizations has been used in many of visualization research prototypes, but the acceptability of such heatmap based visualization has not been studied in the context of geographic information retrieval for lay users.

Most of the users who participated in the preliminary study were constantly involved in the design and assessment of the interfaces described in this work, still the comparative study of proposed interfaces is essential. In future we plan the

extensive qualitative and quantitative evaluation of these interfaces, to conclude the acceptability of these visualizations with regard to user-centered aspects such as exploration ability, information overload and cognitive demand. The methodology discussed by Longo et al. [15,14] would be a guideline to evaluate the proposed interfaces, and to judge the human mental workload in web-design imposed by different interfaces. We would also like to investigate more advanced interaction methods to enhance the usability of proposed visualizations.

Acknowledgments. The authors are grateful to the DFG SPP priority program[6] which funds the project UrbanExplorer.

References

1. Ahlers, D., Boll, S.: Location-based Web search. In: Scharl, A., Tochtermann, K. (eds.) The Geospatial Web. Springer (2007)
2. Ahlers, D., Boll, S.: Retrieving Address-based Locations from the Web. In: GIR 2008: Proceedings of the 5th International Workshop on Geographic Information Retrieval. ACM, New York (2008)
3. Ahlers, D., Boll, S.: Adaptive Geospatially Focused Crawling. In: CIKM 2009. ACM (2009)
4. Aurenhammer, F.: Voronoi diagrams: a survey of a fundamental geometric data structure. ACM Comput. Surv. 23(3), 345–405 (1991)
5. Dykes, J., MacEachren, A.M., Kraak, M.-J.: Exploring Geovisualization. International Cartographic Association Series. Elsevier (2005)
6. Fisher, D.: Hotmap: Looking at geographic attention. IEEE Transactions on Visualization and Computer Graphics 13(6), 1184–1191 (2007)
7. Greene, R., Devillers, R., Luther, J.E., Eddy, B.G.: GIS-based multiple-criteria decision analysis. Geography Compass 5(6) (2011)
8. Harrower, M.A., Brewer, C.A.: Colorbrewer.org: An online tool for selecting color schemes for maps. The Cartographic Journal 40(1) (2003)
9. Holzinger, A.: Usability engineering methods for software developers. Communications of the ACM 48(1), 71–74 (2005)
10. Keim, D.A., Kohlhammer, J., Ellis, G., Mansmann, F.: Mastering The Information Age - Solving Problems with Visual Analytics. Eurographics (2010)
11. Kumar, C.: Relevance and ranking in geographic information retrieval. In: Proceedings of the Fourth BCS-IRSG Conference on Future Directions in Information Access, pp. 2–7. British Computer Society (2011)
12. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical studies in information visualization: Seven scenarios. IEEE Transactions on Visualization and Computer Graphics 18, 1520–1536 (2012)
13. Lloyd, D.: Evaluating human-centered approaches for geovisualization. PhD thesis, City University London (September 2009)
14. Longo, L., Kane, B.: A novel methodology for evaluating user interfaces in health care. In: 2011 24th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6. IEEE (2011)

[6] `http://www.visualanalytics.de`

15. Longo, L., Rusconi, F., Noce, L., Barrett, S.: The importance of human mental workload in web design. In: WEBIST, pp. 403–409 (2012)
16. MacEachren, A.M., DiBiase, D.: Animated maps of aggregate data: Conceptual and practical problems. CaGIS 18(4) (1991)
17. Maguire, M.: Methods to support human-centred design. Int. J. Hum.-Comput. Stud. 55(4), 587–634 (2001)
18. Makropoulos, C., Butler, D.: Spatial ordered weighted averaging: incorporating spatially variable attitude towards risk in spatial multi-criteria decision-making. Environmental Modelling & Software 21(1), 69–84 (2006)
19. Moere, A.V., Tomitsch, M., Wimmer, C., Christoph, B., Grechenig, T.: Evaluating the effect of style in information visualizations. IEEE Transactions on Visualization and Computer Graphics 18, 2739–2748 (2012)
20. Nivala, A.-M., Sarjakoski, L.T., Sarjakoski, T.: User-centred design and development of a mobile map service. In: Hauska, H., Tveite, H. (eds.) Scandinavian Research Conference on Geographical Information Science, Stockholm, Sweden, June 13-15. Proceedings of the ScanGIS, pp. 109–123. ScanGIS' 2005 (2005)
21. Rinner, C., Heppleston, A.: The spatial dimensions of multi-criteria evaluation – case study of a home buyer's spatial decision support system. In: Raubal, M., Miller, H.J., Frank, A.U., Goodchild, M.F. (eds.) GIScience 2006. LNCS, vol. 4197, pp. 338–352. Springer, Heidelberg (2006)
22. Rinner, C., Raubal, M.: Personalized multi-criteria decision strategies in location-based decision support. JGIS 10 (2004)

Immersive Interactive Information Mining with Application to Earth Observation Data Retrieval

Mohammadreza Babaee[1], Gerhard Rigoll[1], and Mihai Datcu[2]

[1] Institute for Human-Machine Communication, Technische Universität München, Munich Aerospace Faculty, Munich, Germany
{reza.babaee,rigoll}@tum.de
[2] Munich Aerospace Faculty, German Aerospace Center, Wessling, Germany
mihai.datcu@dlr.de

Abstract. The exponentially increasing amount of Earth Observation (EO) data requires novel approaches for data mining and exploration. Visual analytic systems have made valuable contribution in understanding the structure of data by providing humans with visual perception of data. However, these systems have limitations in dealing with large-scale high-dimensional data. For instance, the limitation in dimension of the display screen prevents visualizing high-dimensional data points. In this paper, we propose a virtual reality based visual analytic system, so called *Immersive Information Mining*, to enable knowledge discovery from the EO archive. In this system, Dimension Reduction (DR) techniques are applied to high-dimensional data to map into a lower-dimensional space to be visualized in an immersive 3D virtual environment. In such a system, users are able to navigate within the data volume to get visual perception. Moreover, they can manipulate the data and provide feedback for other processing steps to improve the performance of data mining system.

Keywords: Immersive visualization, Information mining, and Dimension reduction.

1 Introduction

The volume of multimedia data in different applications is growing intensively since the last decade. For instance, the amount of collected EO images is increasing dramatically in the order of hundreds of terabytes a day. Simultaneously, new approaches for information retrieval and knowledge discovery are in high demand. Traditional methods for exploring EO data are based on Image Information Mining whose main steps are feature extraction, data reduction, and labeling. Since all these methods are developed to perform automatically, the gap of human interaction is very large. Therefore, developing a new process chain, mainly based on involving humans' perception of data can be a promising solution to fill in the gap. In order to process image data, first the features of the images are represented by discrete feature vectors (e.g., SIFT [1], Weber [2], Color Histogram [3], etc). In order to avoid loss of information during

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 376–386, 2013.

discretization, the dimension of the feature vectors is set rather high. However, high-dimensional features make data understanding and knowledge discovery more challenging.

Visual analytic systems have shown a great contribution in data analysis by providing human with a visual representation of data. However, they all have limitations in the number of data points and dimensionality of feature space [4]. For example, dealing with high dimensionality is the main challenge in the visualization part of these systems due to the limitation in dimensionality of display screens.

In this paper, we propose an Immersive Information Mining framework as a novel visual analytic system to handle large-scale high-dimensional EO images. The main features of this system are: 1) reducing the dimension of high-dimensional data to three dimensions, utilizing a library of state-of-the-art dimensionality reduction (DR) techniques; 2) visualizing the data in an immersive 3D virtual environment. The proposed system allows users to play around with various DR techniques along with different parameters. They can visually compare the visualizations and choose the one that shows the structure of the given feature space better. Further, this system allows users to interact with data. More precisely, the user can change the structure of feature space by changing the position of feature points which can be used not only as a hint for feature descriptors to distinguish various classes better but also to correct available image annotations.

The rest of the paper is organized as follows. In Section 2, we discuss several current visual analytic systems. We illustrate the concept of Immersive Information Mining in Section 3. Section 4 presents samples of immersive visualization of EO images. The evaluation part of system is presented in Section 5, and finally Section 6 presents the conclusion and future works.

2 Related Work

In this section, we briefly review several currently available visual analytic systems. IN-SPIRE [5] is a well-known visual analytic system for document processing comprising, mainly, dimension reduction and clustering. It first extracts high-dimensional features from documents utilizing a bag-of-words model and then applies k-means clustering (with pre-defined number of clusters) on the features for data reduction. In order to visualize features, PCA reduces the dimension of features to two dimensions and then the results are plotted on screen. Another visual analytic system for document processing is Jigsaw [6] using named entities for visualization. In this system, clustering is carried out by the k-means algorithm and the results are plotted on screen. iPCA [7] also applies PCA on high-dimensional data for dimension reduction. Additionally, it visualizes both low-dimensional data along with the principal axes in high-dimensional space via parallel coordinates. Finally, Testbed [4] claims to offer an interactive visual system for dimension reduction and clustering. This system has a built-in library of dimension reduction and clustering techniques. This system aims to help the

user to understand data by visualizing the results of different dimension reduction and clustering methods. It claims to reveal valuable knowledge from data and assist the user to choose the most appropriate processing path along with proper parameter(s).

Since the last decade, numerous projects have utilized virtual reality for information visualization. For instance, VRMiner is a 3D interactive tool for visualizing multimedia data utilizing virtual reality [8]. In this system, a set of numeric and symbolic attributes along with multimedia data (e.g., music, video, and websites) are presented in a 3D virtual environment. But images and videos are displayed on a second PC in order to have a real-time system. Another sample illustrating the usage of virtual reality for information visualization is an application named 3D MARS [9]. This application is mainly for content-base image retrieval, in which the user browses and queries images in an immersive 3D virtual environment. The main aim of this system is visualizing the results in 3D space.

Beside aforementioned technologies for the visualization and exploration of data, Human-Computer Interaction (HCI) has shown valuable contribution in the domain of Data Mining and Knowledge Discovery. The main aim is providing the user with a way to learn how to analyze the data in order to get knowledge to make proper decisions. For example, Holzinger [10] has investigated HCI for interactive visualization of biomedical data. As another example, Wong et al [11] have shown first the similarity between intelligent information analysis and medical diagnosis and then proposed what kind of issues should be considered during the design of an interactive information visualization supporting intelligent information analysis.

As is clear from above, the main processing step in every visual analytic system is dimension reduction. Since the last decade, numerous linear and nonlinear DR techniques have been proposed in different research areas. While linear approaches assume data lies in a linear d-dimensional subspace of a high-dimensional feature space, nonlinear approaches consider data as a d-dimensional manifold embedded in high-dimensional space. The most useful techniques in our project are explained briefly here. Perhaps, the most famous linear algorithm is Principal Component Analysis (PCA) which projects data into d eigenvectors corresponding to d largest eigenvalues of the covariance matrix of the data. Among nonlinear methods, Locally Linear Embedding (LLE) [12] aims to preserve the structure of data during dimension reduction. It assumes that the data belongs to a low-dimensional smooth and nonlinear manifold which is embedded in a high-dimensional space. Then, the data points are mapped to lower-dimensional space in such a way that neighborhood is preserved.

Laplacian Eigenmaps (LE) [13] is a nonlinear technique in the domain of spectral decomposition methods and locally transforms data into low-dimensional space. It performs this transformation by building a neighborhood graph from the given data whose nodes represent data points and edges depict the proximity of neighboring points. This graph approximates the low-dimensional manifold embedded in a high-dimensional space. The eigen-functions of the Laplace

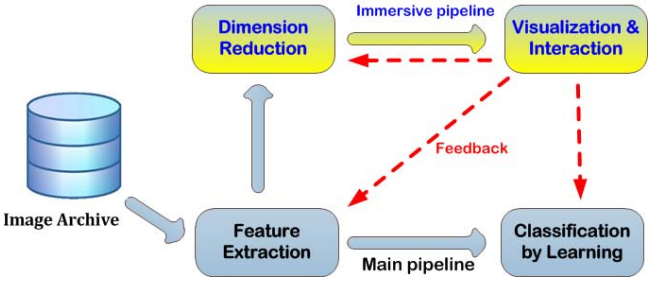


Fig. 1. The workflow of immersive information mining. This comprises two parallel process pipeline. The blue one is the main processing line existing in traditional data mining techniques. The yellow one is our proposed immersive pipeline composed of dimension reduction and visualization.

Beltrami operator on the manifold serve as the embedding dimensions. Stochastic Neighbor Embedding (SNE) [14] is a probabilistic based approach attempting to preserve the neighborhoods of points based on converting the distances into probabilities. Therefor, the neighborhood relation between two data points is represented by a probability such that closer points to a specific point have larger probability than further points. Then, data points are mapped to low-dimensional space such that the computed probabilities are preserved. This is done by minimizing the sum of the Kullback-Leibler divergences of the probabilities.

3 Immersive Information Mining

Our proposed approach for data mining (i.e. Immersive Information Mining) suggests a new processing pipeline comprising feature extraction, dimension reduction and immersive visualization and interaction. This line runs in parallel to the main process pipeline (feature extraction and machine learning) in order to provide users with a mechanism to give feedback to other processing steps. Fig. 1 depicts the diagram of our Immersive Information Mining system.

As Fig. 1 shows, the immersive pipeline consists of two main steps, dimension reduction and visualization. The idea behind these steps is to reduce the dimensionality of the given high-dimensional feature vectors to 3D and then represent the resulting feature vectors in an immersive 3D virtual environment. In our proposed system, the dimension reduction step is a collection of linear and non-linear methods providing a meaningful and compact 3D representation of the feature vectors. Moreover, for visualization and user interaction, we use *Virtual Reality (VR)* technology (we discuss each step in more details in the following sections. See Fig. 2).

3.1 Dimension Reduction for Visualization

Dealing with visual data, representing images by their important features is a vital pre-processing step. Varieties of feature descriptors have been introduced

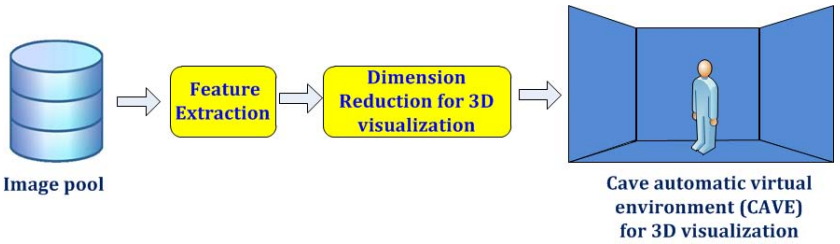


Fig. 2. An immersive visualization system provides the user with a visual representation of data. Here, high-dimensional features are extracted from a database of Earth Observation images and are fed into a dimension reduction technique to be visualized in an immersive 3D virtual environment.

during recent years to describe images from different aspects (e.g., shape, texture, color, etc). In this paper we deal with three different feature descriptors: Scale Invariant Feature Transform (SIFT) [1], Webers Local Descriptors (WLD) [15], and Color Histogram [3]. These descriptors are explored because they represent three different visual contents of images (shape, texture, and color, respectively). In order to cover all the important features of the images, the extracted feature vectors are usually high-dimensional. However, to represent the extracted feature vectors in our virtual environment, we have to reduce the dimension of features to 3D. The DR step of our system currently consists of a number of both linear and nonlinear dimension reduction techniques such as PCA [16], LDA [17], LLE [12], LE [13], SNE [14], and Non-Negative Matrix Factorization (NMF) [18,19]. However, our system allows employing other techniques in addition to the available methods. Providing various DR techniques by the system allows users to switch between different techniques to see how they affect the structure of the feature space during the dimension reduction process.

3.2 Visualization and Interaction

In our visual analytic system, we utilize an immersive 3D virtual environment for visualization of feature vectors and user interaction. To provide a virtual reality environment, the so called Cave Automatic Virtual Environment (CAVE) is used. CAVE consists of four room-sized walls which are intended to play the role of four display screens. They are aligned in such a way as to form a cube-shape space. This configuration allows users to have a 180 degree horizontal view. The computer generated scene is projected onto the walls, using two projectors per wall, in order to have stereoscopic scenarios. Furthermore, a real-time tracking system comprising six infrared cameras, mounted on top of the walls, computes the position and orientation of objects (e.g., Wii controller and glasses) inside the cube. This object tracing is based on the position and orientation of a number of markers attached to every object. The computation behind this projection and tracking is done by a set of PCs in three layers. The layers collaborate in such a way as to provide users a 3D perception of the processed scene. As can be seen

in Fig. 3, the first layer is responsible for processing the signals coming from the user's motion and navigation. In our work, mainly the position and rotation of the head of the user is tracked by tracking the markers mounted on the shutter glasses worn by the user. These glasses help users to have a 3D perception of the scene generated by multiple projections in the cube. Furthermore, as the navigation tool (wand), a Wii controller in this layer, provides navigation as well as scene management signals. The middle layer, which consists of a master PC, is responsible for receiving signals and commands from the first layer. The master PC in this layer modifies the scene interpretation based on the incoming signals and then sends rendering signals to the third layer. The third layer comprises four synchronized PCs for rendering and displaying the scene on the walls. It is possible to have a strong PC for rendering instead of a cluster of 4 PCs, but such a system should be specially powerful which is much more expensive than a cluster of normal PCs. Finally, all PCs are connected to a wired LAN and are synchronized in order to have a real-time visualization. We did not use wireless LAN due to two main reasons. First, all PCs are close to each other, and second, wireless LAN has higher latency effecting the real-time property of system. The software library used for rendering the 3D environment is named 3DVIA Virtools which is used to create 3D real-time applications.

4 Experiments

In order to show how our visual analytic system performs for different kinds of data, we visualize three different data sets containing multi-spectral and optical EO images (UCMerced-LandUse dataset [20], optical multimedia images (Corel dataset [21]), and Synthetic Aperture Radar (SAR) images. The UCMerced-Land-Use dataset is an annotated dataset consisting of 21 classes where each class contains 100 image patches from aerial orthography. The Corel dataset comprises 1500 multimedia images categorized in 15 different groups where each group contains 100 images. Finally, the data set of SAR data is an annotated collection of 10000 SAR images. These images are categorized in 100 classes with 100 images in each. Samples of these data sets are shown in Fig. 4.

In order to process the aforementioned data sets, the images are represented by three different feature descriptors; namely, SIFT [1], WLD [15], and color-histogram [3]. These descriptors are high-dimensional feature vectors whose dimensionality is reduced in the DR step of our system to be visualized in the virtual reality space. In our experiments, we apply three different dimensionality reduction techniques to the given feature vectors to reduce their dimensionality to 3D (e.g., LE [13], SNE [14], and LLE [12]). Fig. 5–7 depict the visualization of our data sets in the CAVE.

In our visualization system, the user is allowed to navigate within the data and select some features by a provided selection tool. Zooming is completely provided and the user has the ability to view features in different views. Additionally, when it is necessary, the system can visualize the images corresponding to feature points.

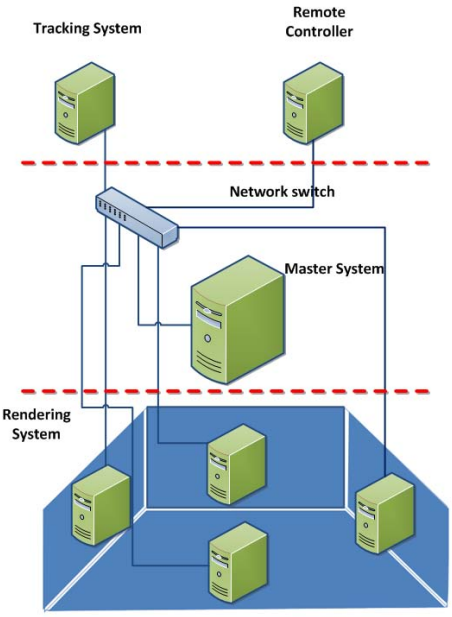


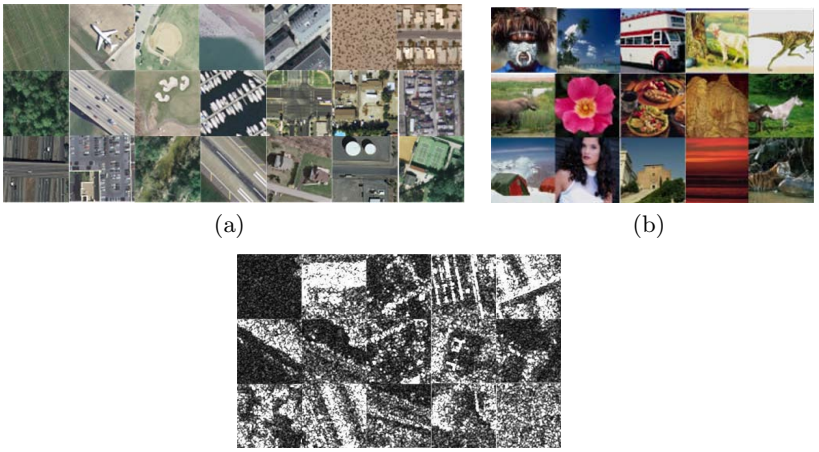
Fig. 3. The physical diagram of immersive visualization. The visualization system is compose of three layers with different responsibilies. The first layer comprises two PCs for motion capturing (tracking) and control. A master PC in the middle layer for the synchronization, and finally four PCs for rendering for each wall of the CAVE. All PCs are connected together via an Ethernet network.

5 Evaluation

Our proposed system comprises, mainly, dimension reduction and interactive visualization. Therefore, the type of evaluation for dimension reduction is different from visualization.

The used techniques for the quality assessment of the dimension reduction step include a local continuity meta-criterion (Qnx) [22,23], trustworthiness and continuity measures (Qtc) [24], and mean relative rank error (Qnv) [25,26]. We extracted Color-Histogram, SIFT, and Weber features from Merced and Corel data sets and then utilized Laplacian Eigenmaps (LE) [13], Stochastic Neighbor Embedding (SNE) [14], and Locally Linear Embedding [12] as dimension reduction techniques to reduce the dimensionality of extracted features. The 9 different combinations of features and dimension reductions, which are called methods here, are; 1) Color-LE, 2) Color-SNE, 3) Color-LE, 4) SIFT-LE, 5) SIFT-SNE, 6) SIFT-LLE, 7) Weber-LE, 8) Weber-SNE, and 9) Weber-LLE. The aforementioned quality measures were applied on these methods whose results are depicted in Fig. 8.

It can be concluded that the performance of features and dimension reduction depends on the type of data set. For instance, the SNE applied on extracted



(a) (b)

(c)

Fig. 4. Sample images from used datasets; a) Corel; b) UCMerced-LandUse; c) Radar Images

(a) (b) (c)

Fig. 5. Three sample images of immersive visualization of UCMerced dataset

(a) (b) (c)

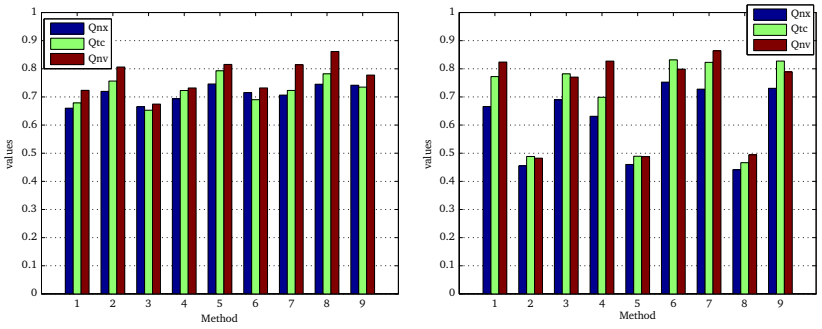
Fig. 6. Three sample images of immersive visualization of Corel dataset

Weber features from Corel data set excels the SNE applied on Weber features from Corel data set.

In order to evaluate the visualization part of proposed approach, we plan to accomplish useability studies and define some objective and subjective measurement. For instance, how much time is needed to accomplish a special task? Or

(a) (b) (c)

Fig. 7. Three sample images of immersive visualization of SAR images



(a) (b)

Fig. 8. The quality assessment of dimension reduction techniques applied to extracted features from Merced and Corel data sets. A combination of three different features (color-histogram, SIFT, and Weber) and three different DR techniques (LE, SNE, LLE) gives 9 feature-DR methods which are: 1) color-LE, 2) color-SNE, 3) color-LE, 4) sift-LE, 5) sift-SNE, 6) sift-LLE, 7) weber-LE, 8) weber-SNE, 9) weber-LLE. Those methods whose quality measurements are closer to 1 have better performance. a) Results from Corel data set; b) Results from Merced data set.

how robust is the data manipulation? or how much time is needed for training. As subjective measurements, data understanding and manipulation are considered. However, it is clear that visualization in the CAVE is superior to the head-mounted-display based and monitor-based visualization since in the CAVE the user is able to walk freely and look at the data from different angles.

6 Conclusion and Future Work

In this paper, we present a virtual reality based visual analytic system, a so-called immersive information mining system with application to earth observation images. The main features of this system are dimension reduction and immersive visualization. Technically, we reduce the dimension of high-dimensional feature vectors to 3D and then visualize them in an immersive 3D virtual environment.

This environment allows the user to navigate inside the data and get a visual understanding of structure of data. The feedback coming from visual analytic helps the user to choose a proper processing path with suitable parameter(s). Another advantage of this approach is interactivity with data. Potentially, the user could be able to manually change the structure of data and impose constraints on the learning process which is considered as future work.

Acknowledgments. This work is supported by PhD scholarship Award by Munich Aerospace Faculty of German Aerospace Center (DLR). The authors would like to thank reviewers for their valuable comments.

References

1. Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
2. Bahmanyar, R., Datcu, M.: Measuring the semantic gap based on a communication channel model (2013)
3. van de Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
4. Choo, J., Lee, H., Liu, Z., Stasko, J., Park, H.: An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data. In: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, pp. 865402–865402 (2013)
5. Wise, J.A.: The ecological approach to text visualization. Journal of the American Society for Information Science 50(13), 1224–1233 (1999)
6. Stasko, J., Görg, C., Liu, Z.: Jigsaw: supporting investigative analysis through interactive visualization. Information Visualization 7(2), 118–132 (2008)
7. Jeong, D.H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., Chang, R.: ipca: An interactive system for pca-based visual analytics, vol. 28, pp. 767–774. Wiley Online Library (2009)
8. Azzag, H., Picarougne, F., Guinot, C., Venturini, G., et al.: Vrminer: A tool for multimedia database mining with virtual reality. In: Processing and Managing Complex Data for Decision Support, pp. 318–339 (2005)
9. Nakazato, M., Huang, T.S.: 3d mars: Immersive virtual reality for content-based image retrieval. In: IEEE International Conference on Multimedia and Expo, vol. 46 (2001)
10. Holzinger, A.: On knowledge discovery and interactive intelligent visualization of biomedical data-challenges in human-computer interaction & biomedical informatics. In: 9th International Joint Conference on e-Business and Telecommunications (ICETE 2012), pp. IS9–IS20 (2012)
11. Wong, B.L.W., Xu, K., Holzinger, A.: Interactive visualization for information analysis in medical diagnosis. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 109–120. Springer, Heidelberg (2011)
12. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)

13. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15(6), 1373–1396 (2003)
14. Hinton, G., Roweis, S.: Stochastic neighbor embedding. Advances in Neural Information Processing Systems 15, 833–840 (2002)
15. Chen, J., Shan, S., Zhao, G., Chen, X., Gao, W., Pietikainen, M.: A robust descriptor based on weber's law. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–7. IEEE (2008)
16. Jolliffe, I.T.: Principal component analysis, vol. 487. Springer, New York (1986)
17. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.: Fisher discriminant analysis with kernels. In: Proceedings of the 1999 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing IX, pp. 41–48. IEEE (1999)
18. Seung, D., Lee, L.: Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems 13, 556–562 (2001)
19. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
20. http://vision.ucmerced.edu/datasets/landuse.html
21. http://vision.stanford.edu/resources_links.html#datasets
22. Chen, L., Buja, A.: Local multidimensional scaling for nonlinear dimension reduction, graph layout and proximity analysis. PhD thesis, Citeseer (2006)
23. Chen, L., Buja, A.: Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. Journal of the American Statistical Association 104(485), 209–219 (2009)
24. Venna, J., Kaski, S.: Local multidimensional scaling. Neural Networks 19(6), 889–899 (2006)
25. Lee, J.A., Verleysen, M.: Nonlinear dimensionality reduction. Springer (2007)
26. Lee, J.A., Verleysen, M.: Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing 72(7), 1431–1443 (2009)

Transfer Learning for Content-Based Recommender Systems Using Tree Matching

Naseem Biadsy[1], Lior Rokach[2], and Armin Shmilovici[2]

[1] Deutsche Telekom Labs at Ben-Gurion University
Beer-Sheva 84105, Israel
naseem@cs.bgu.ac.il
[2] Department of Information System Engineering Ben-Gurion University
Beer-Sheva 84105, Israel
{liorrk,armin}@bgu.ac.il

Abstract. In this paper we present a new approach to content-based transfer learning for solving the data sparsity problem in cases when the users' preferences in the target domain are either scarce or unavailable, but the necessary information for the preferences exists in another domain. Training a system to use such information across domains is shown to produce better performance. Specifically, we represent users' behavior patterns based on topological graph structures. Each behavior pattern represents the behavior of a set of users, when the users' behavior is defined as the items they rated and the items' rating values. In the next step, a correlation is found between behavior patterns in the source domain and target domain. This mapping is considered a bridge between the two. Based on the correlation and content-attributes of the items, a machine learning model is trained to predict users' ratings in the target domain. When our approach is compared to the popularity approach and KNN-cross-domain on a real world dataset, the results show that our approach outperforms both methods on an average of 83%.

Keywords: Recommender-Systems, Transfer Learning, Content-based, Behavior Patterns.

1 Introduction

Cross domain recommenders [2] aim to improve recommendation in one domain (hereafter: the target) based on knowledge in another domain (hereafter: the source). There are two approaches for cross domain: 1) those that use a mediator to construct and initialize an empty user model, or that enrich the existing user model by using provided data or a partial user model in another domain or service; 2) those that do not use a mediator. In this paper we propose a content-based cross domain RS that uses a mediator to suggest items to **a new user** in a target domain who has only rated items in the source domain. We assume that: 1) the target domain has very high sparsity; 2) the source domain has low sparsity; and 3) the new user has already rated some items in the source domain. Based on these assumptions, traditional recommendation algorithms do

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 387–399, 2013.

not work well in the target domain for two reasons: first, high sparsity; second, the new user and cold-start issues.

The problem: Given a set of items T in a target domain, and given a user u who has already rated certain items in the source domain but not in the target, what are the N most preferred items in T by user u?

We begin solving this problem by presenting a new graph-based transfer learning method for content-base recommendation. The concept of behavior tree is defined as a topological representation of users' behavior patterns. By ***users' behavior*** refers to the items rated by a user and the rating values the user has assigned to these items. Based on these trees, we find correlated behavior patterns in the source and target domains which are considered bridges between the two domains. Then, the data of each bridge is combined with the content data of the items in a one features vector. This vector is used as a sample in the training data for a machine learning algorithm. Here common users are assumed to exist between the source and target domains.

The paper's innovation is (1) the graph structure employed for dealing with the problem, instead of just using raw numbers. The trees structure provides a rich representation of the users' behavior. First, the structure enables items to be clustered according to the users' behavior, and the clusters to be represented by a forest of behavior trees. Second, the tree's structure defines a hierarchy among the items, this hierarchy is important because it shows how much the item represents the behavior pattern it belongs to. (2) Using the items' content data from a rich real world dataset with transfer learning recommenders that are employed to improve classifier performance. Finally, we find a correlation (mapping) between items in the source domain and items in the target domain, so the mapping is obtained by running a process of tree matching between trees in the source and trees in the target. Furthermore, from the content data, more features are gained to represent the training samples for building the recommendation model.

2 Related Work

Several recent studies apply cross domain techniques in order to achieve performance improvement over a single domain recommender [9]. One of the first studies [13] is a user study that examined the fundamental assumption that users share similar preferences across domains. The first case study presented in this work tested the correlations between related items in different domains. Their similarity across domain was determined according features or topical similarity. For example, "Harry Potter" the movie and the game are considered similar items, as are Andy Lau the singer and the song sung by her. This study found that positive correlations of user preferences usually exist across domains. For instance, users having positive evaluation of a person as an actor/actress also exhibit a strong interest in the same person as a singer.

The second case study examined the correlation in users' preferences on items of the same genre but of different domains (e.g. preferences of action movies and

action games). The results showed a positive correlation of users' preferences of the same genres across domains. The results reinforced the assumption that user interest exists across domains.

The study's evaluation was based on user's questionnaires (the users were asked for their preferences) and not on an actual recording of system usage, with a limited number of items and users. Moreover, it was based on the aggregate ratings that users provided rather than on an individual's ratings. Therefore, we can only infer the existence of general trends towards preferences across domains. However, this is the only user study that has attempted to examine the assumption that all other studies assume about the relatedness of preferences across domains.

Among the studies dealing with cross domain recommenders, one can distinguish between two main approaches for applying data from different domains are discernible. One aims at constructing a unified user model from the different domains, while the second approach uses "Transfer Learning" techniques. The following sections review the characteristics of the two approaches.

User-Model Mediation. Perhaps the simplest approach to cross domain recommendations is to import relevant data from a different domain and aggregate it with the original target data. The aggregation is simple when the domains share information on the same users. Berkovsky et al. [3] refer to this problem as cross user-model mediation and suggest the following definition: A user model is the data that the RS has collected on the user in order to provide him with personalized services. Most of the recommender systems maintain UMs that are tailored to their specific application or domain. According to Berkovsy et al. [1], since the quality of the personalized services depends heavily on the UM's characteristics and accuracy, different services would benefit from enriching their UMs by importing, translating and aggregating partial UMs from related domains or services.

Several UM mediation techniques are represented in [4] and [3]. The target RS sends a request for rating prediction; the other RSs in the different domains send information back to the target system; and each suggested technique differs in the information returned to the target system:

Centralized Prediction - Each RS sends back its local users-items-ratings matrix. The target RS unifies all the given matrices with its own local matrix, and run collaborative Filtering (CF) on the unified matrix. Distributed Peer Identification - this technique assumes that if two users are similar in some domain, they may also be similar in another related domain. In practical terms, each RS computes the k-nearest neighbors of the given user-id according to its local matrix and returns their identities. The target RS unifies the given answers with its local nearest neighbors and computes the rating using CF. Distributed Neighborhood Formation - Each RS computes the similarity between the active user i and the other users using their own local rating matrix. A set of K nearest-neighbors is selected, and their identifiers, together with their similarity are sent to the target RS. The target RS averages the domain-specific similarity values into the overall similarity metric using inter-domain correlation values. When

the overall similarity value is computed, K nearest neighbors can be selected and the predictions generated. Distributed Prediction- Each RS that contains rating data on the desired item computes locally the rating using CF, and sends back the rating to the target. The target RS calculates its local CF rating, and calculates the average of all the ratings.

Experiments were conducted to compare these approaches to a baseline approach that used only the target domain data. Results showed that in some conditions integrating neighbors (2nd approach) and integrating recommendation (3rd approach) can be beneficial, as they may improve, respectively, the coverage and accuracy of the generated recommendations. However, in other cases, when the target system has enough rating records, it is preferable to use only the target data since the remote systems data may be interpreted as noise.

The study has certain limitations. The method assumes that the different domains share the same users who are identifiable in different applications. Yet, in many applications no such overlapping exists and even if there are common users, they are unidentifiable. In addition, some of the suggested approaches, for example the Distributed Prediction approach, assume that at least a large portion of the domains share the same items with the target domain, which means that such methods cannot be applied to totally different domains like movies and songs.

A similar approach is described in [6]. The authors generated a uniform UM approach that aggregated features from different domains, and mapped them to relevant domains.

In conclusion, the main advantage of the user-model mediation is by its simplicity and intuitive implementation. However, it can be applied only to particular problems when the different sources share implicit resources like users, items or features. Since one of our goals in this work has been to avoid restricting users or items that overlap between domains, we did not use this approach.

Transfer Learning. Transfer learning (TL) is a relatively new area of research (since 1995) in machine learning whose aim is to extract knowledge that was learned for one task in a domain and use it for a target task in a different domain [11]. In the field of machine learning we usually train a model based on available data for the problem that we are interested in (training data). The model is used for predicting behavior in examined domain (using the testing data). For example, in recommender systems the model can help us predict whether or not the user will like a movie.

Transfer Learning is a very intuitive notion since human nature is to transfer learning from different aspects of life. People can intelligently apply previously learned knowledge to solve new problems faster or better. For example, when learning a foreign language such as French, knowledge of Romanian can be used to pick up new words and skills that we acquired in the past can be employed to assimilate the new language. Another example, when learning a new programming language like C++, our knowledge of Java and our general programming

skills can be exploited to our benefit. TL also enables us to retain and reuse previously learned knowledge [11], and this may reduce the need and effort to recollect training data in new applications.

An important TL characteristic is that it does not always require content overlap between the different domains. Regarding the recommender systems application, TL does not require that the source and target domain share users or items, since it aims at finding common consumption patterns that exist in related domains by recognizing latent behavior groups [10]. Consider, for example, the music and games domains. Although they do not appear strongly connected, the same latent groups of users are still found in both domains. For example, there might be a group of consumers that always purchase new trendy items, or another group which likes to consume low cost products, and still other users who especially enjoy children's items.

Transferring knowledge between domains is a daunting task because the knowledge of one domain is not guaranteed to be useful for another domain. The success of transfer learning depends on a variety of factors, e.g. the degree of domain correlation (for example, are movies and books more closely correlation than movies and jokes), the data characteristics (sparsity level, rating scale etc.), and whether the domains share common resources such as items or users.

There are only a few transfer learning applications for the recommender system. Most of the works that used TL for recommender systems are based on collaborative filtering. Next several works are reviewed that employ TL for recommender systems and discuss their limitations and difference from our work.

Li et al [10] may be the most relevant work. They suggested that when insufficient data in the target domain prevents training an accurate recommendation model, useful knowledge can be borrowed from a different domain and its data used to train the model. They introduced the idea that rating matrices of different domains may share similar user-item rating patterns. Thus, they adopted a user-item rating matrix of the source domain, referred to as a "codebook", and transferred the rating patterns to the target domain in order to fill in the target domain's missing values. Their algorithm consists of two steps. First, a rating pattern (codebook) of the dense domain is created, which summarizes the original rating data. Second, the codebook is expanded in order to learn the missing values in the target domain.

3 The Recommendation Framework

Our framework deals with the abovementioned problem in three phases: The first is described in (3.1) and aims to preprocess the users' data in the source and target domains. Preprocessing consists of three steps: 1) building behavior trees for each domain; 2) tree matching; 3) building training samples. The second phase (3.2) is for training a model on the training set. The third phase, described in (3.3), is for recommending items to new users based on their behavior in the source domain.

3.1 Preparing Training Data

Building Behavior Graphs. Constructing the graphs requires four steps. **First**, the same item with different rating values is separated by expanding each item into k items, where k equals the number of possible rating values in the domain. For example, if item i was rated as $r1$ by one group of users and the same item received an $r2$ rating from a different group of users, we consider them as different items and represent them by i_{r1} and i_{r2}. Note: the number of items after this step is $k \times number\ of\ origin\ items$. **Second step.** The separated items are sorted by their popularity, that is, by the number of the user ratings. Table (1) is an example of a rating matrix, where the popularity of item 1_2 (item 1 with rating 2) in this matrix is 3.

Table 1. Rating matrix, possible ratings [1, 2, 3] (k=3)

	item1	**item2**	**item3**	**item4**	**item5**
user1	2			1	1
user2	1		1	1	3
user3	2	1	2	3	
user4	3		2		1
user5		3	3	1	1
user6	1	3		1	
user7	1		1	2	3
user8	2	3	3	2	1
user9	1	3		1	
user10	1		1		3

Third step: A topological representation is adopted by taking the sorted items set of each domain and representing them by a disconnected weighted graph. The graphs are constructed as follows: **Nodes**: Each item on the sorted items list is represented by one node (note: the node represents a pair of an item and one rating value). **Edges**: An edge is found between two nodes if there are common users exist who rated both items represented by both nodes (note: rating must be the same as in the node). **Weight**: The weight of the edge between two nodes is defined as the Jaccard coefficient between the sets of users who rated the items represented by the nodes. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets (Wikipedia):

Similarity(A, B) = J(A, B) $= \frac{|A \bigcap B|}{|A \bigcup B|}$. If the two sets A and B are empty then we define $J(A, B) = 0$. The max value of this similarity when A, B are finite is 1, and the minimal value is 0.

If the weight of edge is less than a given threshold [1] we drop the edge. Note, a one-to-one mapping exists between the nodes and the items, thus the term node

[1] This threshold depends on different parameters, such as the number of users in the domain, the number of items, the number of ratings, etc. We set it as 0.5 in our experiments.

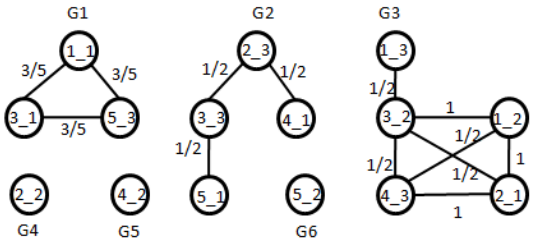


Fig. 1. Behavior graphs (disconnected) of table (1) when threshold = 0.5

or item can be used to refer to the same element. Figure (1) shows the behavior graphs (disconnected) obtained from table (1) when threshold = 0.5.

Fourth step: Each component in the disconnected graph is converted into a topological tree which we refer to as *a behavior tree*. This is another representation of the users' behavior in the graph based on the graph's structure. Transferring a graph into a tree is a known problem. For example, in [5] the authors showed a general transformation between graphs, trees and generalized trees by introducing a special transformation that uses the DIJKSTRA distance between nodes to define a multi-level function which assigns each node of the graph a level value that introduces a hierarchy.

This process takes place by adding an artificial node as the root for each tree, and connecting all the popular [2] nodes in the graph to the node. The weight of each child in level one is equal to the child's popularity. Then a recursive greedy algorithm is used to complete the tree by adding all the other nodes on the graph to it. This is done by connecting each child (who is not in the tree yet) to the parent who is connected to the highest edge weight in the original graph.

The weight of each node (except in first level) in the tree is the same as the edge from the parent to the child in the graph. The idea behind adding an artificial node is to make sure that all the popular items in the graph appear in the first level of the tree and the less popular items appear in deeper levels.

The motivation for moving from graphs to trees is the tree's simpler hierarchical structure than that of graphs, and the tree's greater ability to represent sufficient data of the behavior graphs. Figure (1) represents the graphs obtained from the rating matrix in Table (1), and Figure (2) illustrates the behavior trees based on the behavior graphs in Figure (1), where T_i is the behavior tree of graph G_i, $i : 1 \rightarrow 6$, and the nodes with labels A1-A6 are artificial.

Tree Matching. The first task is to decide for each tree in the source domain which tree in the target domain is the best for matching. Tree matching appears in various fields such as image clustering, RNA, chemical structure analysis, and so fourth. Thus, many method were defined to measure similarity between

[2] The most popular items in each graph are those which have greater popularity than the average popularities in the same graph.

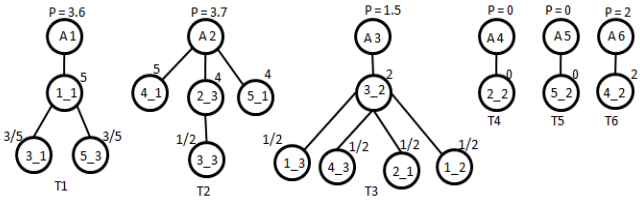


Fig. 2. Constructed behavior trees based on the behavior graphs (1). "P" is the popularity threshold.

trees, for instance largest or smallest common subtree, tree edit distance, or the transferable ratio between two trees [8].

Since we have an advantaged by the common users, the similarity between two trees is defined the same way that similarity between two items is defined in (3.1), but here it is referred to on a set of users instead of items:
Tree-Similarity(T1, T2) $= J(U(T1), U(T2)) = \frac{|U(T1) \bigcap U(T2)|}{|U(T1) \bigcup U(T2)|}$
$U(T) = \bigcup_{i=1}^{|T|} (u \mid \text{user } u \text{ rated item } i \text{ in } T)$.

The similarity measure relates to the content of the trees instead of the structure, so if two trees have the same set of users then they receive a higher similarity value than trees that do not have common users. However, after a tree in the target has been found for each tree in the source that it should be matched with, we run the tree matching process to match the nodes in both trees, and obtain the advantages of the structure of the trees as well. Eventually, a set of pairs of matched nodes is arrived at that are the bridges between the items in the source and the items in the target.

Building the Training Samples. This is the final step in preparing training data. Here the training set must be prepared for a supervised machine learning task. A special structure for features vectors is found that combines the data of users' behavior in the source and target domains, and the items content data. This is done by taking all the matched nodes from the previous step. Then, for each two matched nodes n_s and n_{tr} that represent item t_s in the source with rate r_s and item t_{tr} in the target with rating r_{tr}, respectively, one features vector is defined as following:
$[f_1(t_s), ..., f_n(t_s), r_s; f_1(t_{tr}), ..., f_m(t_{tr}), r_{tr}]$, where $f_i(t)$ is the value of feature i for item t. This structure is meaningful for cross-recommendation because it provides a full image of the user's behavior from the previous step and the features of the items they rated, since it combines the features with the rating values for each group of users, For example, if the users who like books with features $[b_1, b_2, ..., b_n]$, like movies with features $[m_1, m_2, ..., m_m]$, then a new user u with similar behavior in the books domain will be in the same behavior tree and probably has similar behavior toward movies with features $[m_1, m_2, ..., m_m]$.

3.2 Model Training

Here a model of a multiclass classifier is built on the training samples that combine features of items in the source, ratings in the source, features of items in the target, and ratings in the target. Since our goal is to predict the user's rating of an item in the target domain, the class of the classifier is the last attribute in the features vector, where all the other attributes in the vector represent the features. This model is expected to predict rating for an item t in the target domain for a new user u that has rated only items in the source domain. The number of the classes equals the number of possible rating values in the target domain.

3.3 Recommendation

The recommendation task is based on the model described in the previous task (3.2). When the system is asked to recommend the top N items in the target domain for a new user u, the system ranks each item tr in the target domain based on the items rated by the user in the source domain. The ranking is carried out by building *a features matrix* for each item tr in the target as follows:

- The number of rows in the matrix equals the number of items the user rated in the source.
- The number of the column in the matrix equals $n + m + 2$, where n is the number of the item's features in the source, m is the number of the item's features in the target domain and + 2 for the rating in the source and the class (rating in target).
- The value in row i, column j, equals to:
  If $j < n + 1$ Then: The value of feature j in item i (in the source)
  If $j = n + 1$ Then: The rating value that the user u rated item i (in the source)
  If $j > n + 1$ Then: The value of feature $j - (n + 1)$ of the item tr (in the target)
  If $j = n + 2$ Then: The value is "?" (missing)

To find the rank of a features matrix M, we first run the classifier on each row in the matrix in order to return the vector of the predicted probability for each class. This vector is called *the distribution vector* and represents the probability for each class, so the value in entry i equals the probability that this sample is classified as i, and the vector's size equals the number of the classes (possible ratings). For each row we take the distribution vector P returned by the classifier and compute the ***expected rating***[3] as follows: ***expected rating*** $= \sum_{j=1}^{k} j \times P[j]$, when k is the size of P. The rank of the features matrix M is defined as the average of the expected ratings of all the rows in M: $Rank(M) = \frac{\sum_{j=1}^{z} ER(j)}{z}$, where $ER(j)$ is the expected rating for row j and z is the number of the rows

[3] Ratings here are natural numbers between *1* and k when *1* is the minimal value, and k is the maximal value.

in M. Then the matrices are sorted by the rank value, and the target items represented by the top N matrices are sent back as the recommended items for the user u.

4 Experiment

In this chapter we investigate whether the additional knowledge gained from the source domain and the content of the items can improve the recommendation in the target domain. Our approach is compared with the popularity, which is a single domain approach generally used for recommendation to new users, and with the KNN-cross-domain approach that uses both domains but does not use the content data.

4.1 Dataset

We used the Loads data-set in our experiments. Loads data is a real-world content dataset that includes different domains such as videos, music, games (common users among the domains). Music-loads was chosen as the source domain and game-loads as the target domain, and 600 common users were extracted from both domains, 817 items from the music domain with 18,552 ratings, and 1264 items from the games domain with 17,640 ratings. This dataset is event-based, thus we manipulated it and converted the events to ratings by weighting the events by their type. For example, the event *user buys an item* that was converted to the max rate value. For this experiment we used a binary rating.

4.2 Evaluation and Metrics

Our goal is to recommend a set of N items that may find the interest of the new user in the target domain. This kind of recommendation refers to recommending good items [7] or *top-N* items. The correct method of evaluating the recommender in this case is by measuring the precision at N (or Top-N precision) which is the number of interesting items from the recommended items [7]. Since we have content data, a recommended item is considered to be a true positive if it is similar to 80% of the positively rated items.

4.3 Baseline

Our method, referred to as BGM (Behavior Graph Matching), is compared with two base-line recommenders:

KNN-cross-domain: This recommender is a cross domain recommender based on collaborative filtering with Pearson's correlation coefficient method. The main idea behind the method is to find the K-nearest-neighbors of the active user in the source domain who also rated items in the target domain, and to consider them as the K-nearest-neighbors of the active user, who is also in the target

domain, and then predict the user's ratings in the target domain based on his/her neighbors' ratings in the target domain.

Popularity: This is a naive method that recommends the popular items to the new users. This method is the simplest to implement and sometimes outperforms other complex algorithms [12] especially in a domain where most of the users have similar preferences.

4.4 BGM versus Popularity

We evaluate the two methods by 10-fold cross validation on the Loads dataset, when each fold includes 60 users for testing and 540 users for training. The test set was considered as the new users' set in the target domain who have ratings just in the source domain. The behavior patterns were based on all the users in the source domain, and users who belong to the training set in the target domain. For each user in the test set, the recommender was asked to recommend the *Top-N* items when *N= 5, 10, 15, 20, 50, and 100*, and for each set of recommended items the top-N precision was measured per user. The popularity algorithm employs the training set in the target domain to find the popular item that was recommended for each of the test users, then the top-N precision was compared for each N, and for each N the average of the top-N precision was found for all of the users.

Figure (3) shows the results, where we divided into 6 groups, each group representing a different N value and containing three columns, the first one representing the average *Top-N* precision value for the BGM, and the second representing the average of *Top-N* precision value for the popularity algorithm. Note that BGM outperforms the popularity in all the cases except when *N= 100*. The reason is BGM tries to recommend items which are similar to the user's behavior, thus, with small values of N BGM knows better than popularity what items are related to the user, but sometimes users behave in a different way than their normal behavior, such as one decide to buy a book because it is a little bit popular and there is a good offer about it. Here, popularity recommender will recommend this item specially when N is big, while BGM will not recommend this item because it is not similar to the user's preferences. Thus, our results-based conclusion: it is better to use the source domain to improve recommendation in the target domain, specially in cases there is a limitation of the number of items to recommend.

4.5 BGM vs. KNN Cross-domain

This experiment was designed to compare the performance of our approach with the collaborative-filtering cross-domain approach which also uses the knowledge in the source domain to return recommendations to the target domain. The main goal was to determine whether the BGM method makes maximum use of the content data of the items, and to check whether the proposed behavior patterns work well for representing the users' behavior and knowledge transfer. As in the first experiment, this on was conducted with a 10-fold cross validation on Loads dataset.

When comparing the results in Figure (3), where the third column in each group represents the value of top-N precision of the KNN-cross-domain, note that BGM outperforms the KNN-cross-domain in 83% of the cases. The only case where KNN-cross-domain outperforms BGM is when *N=100*. This is because of a similar reason to the one described above: sometimes people make exceptions in their behavior, such as one decide to go to watch a movie just to join friends, so BGM will not perform well with these cases, however, people usually have a consistence behavior, and recommendation is done with small values of *N*.

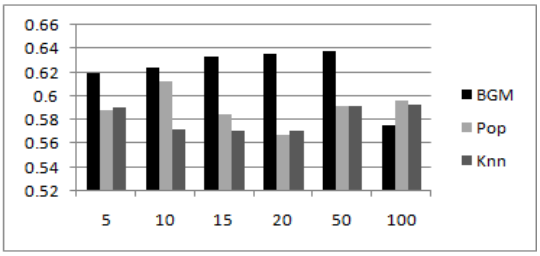


Fig. 3. Top-N precision values for BGM, Popularity, and Knn-cross-domain, with different N values

4.6 Statistical Analysis

Paired t-tests were performed on the same results that were received by running 10-fold cross-validation with BGM, Popularity and KNN-cross-domain on Loads dataset, and their statistically significant was checked. Each user was considered a participant, and each of the recommenders was considered a different method of the test.

A table of 600 users (rows) was obtained, and for each user the top-N precision value was received when *TopN*= 5, 10, 15, and 20 with each of the methods (12 columns per user). Then the t-test was performed for every two relevant columns (with the same N) . All the results are statistically significant, which means that our method performance is decidedly better than the other two when the number of recommended items is 5, 10, 15, and 20, and the size of the sample is 600 participants.

5 Conclusions

Our paper has presented and evaluated a novel method of transfer learning in content-based recommenders by using a new topological structure that we call a behavior graph. The main idea of using such structure is its ability to represent rich data about the users' behavior, and the relation between items in a structure. By employing tree matching methods we discovered a correlation between items in the source and items in the target.

We compared our method with the popularity approach, which is generally used with recommendations to new users, and with the KNN-cross-domain method. The comparison was based on a real-world dataset called Loads dataset, and the Top-N precision metric was evaluated by 10-fold cross validation.

The results show that our method (referred to as BGM) outperforms the popularity and KNN-cross-domain methods in the majority of cases. Our conclusion: it is preferable to use the data in the source domain and the item's content data when dealing with this kind of recommendation problem.

References

1. Berkovsky, S., Kuflik, T., Ricci, F.: Cross-technique mediation of user models. In: Wade, V.P., Ashman, H., Smyth, B. (eds.) AH 2006. LNCS, vol. 4018, pp. 21–30. Springer, Heidelberg (2006)
2. Berkovsky, S., Kuflik, T., Ricci, F.: Entertainment personalization mechanism through cross-domain user modeling. In: Maybury, M., Stock, O., Wahlster, W. (eds.) INTETAIN 2005. LNCS (LNAI), vol. 3814, pp. 215–219. Springer, Heidelberg (2005)
3. Berkovsky, S., Kuflik, T., Ricci, F.: Cross-domain mediation in collaborative filtering. In: Conati, C., McCoy, K., Paliouras, G. (eds.) UM 2007. LNCS (LNAI), vol. 4511, pp. 355–359. Springer, Heidelberg (2007)
4. Berkovsky, S., Kuflik, T., Ricci, F.: Distributed collaborative filtering with domain specialization. In: Proceedings of the 2007 ACM Conference on Recommender Systems, pp. 33–40. ACM (2007)
5. Emmert-Streib, F., Dehmer, M.: Topological mappings between graphs, trees and generalized trees. Applied Mathematics and Computation 186(2), 1326–1333 (2007)
6. González, G., López, B., de la Rosa, J.L.: A multi-agent smart user model for cross-domain recommender systems. In: Proceedings of Beyond Personalization (2005)
7. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22, 5–53 (2004)
8. Jiang, T., Wang, L., Zhang, K.: Alignment of trees - an alternative to tree edit. In: Crochemore, M., Gusfield, D. (eds.) CPM 1994. LNCS, vol. 807, pp. 75–86. Springer, Heidelberg (1994)
9. Li, B.: Cross-domain collaborative filtering: A brief survey. In: 2011 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1085–1086. IEEE (2011)
10. Li, B., Yang, Q., Xue, X.: Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction. In: Proceedings of the 21st International Joint Conference on Artifical Intelligence, pp. 2052–2057. Morgan Kaufmann Publishers Inc. (2009)
11. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
12. Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.M.: CROC: A New Evaluation Criterion for Recommender Systems
13. Winoto, P., Tang, T.: If you like the devil wears prada the book, will you also enjoy the devil wears prada the movie? a study of cross-domain recommendations. New Generation Computing 26(3), 209–225 (2008)

Mobile Movie Recommendations with Linked Data

Vito Claudio Ostuni, Giosia Gentile, Tommaso Di Noia, Roberto Mirizzi, Davide Romito, and Eugenio Di Sciascio

Polytechnic University of Bari, Italy
{ostuni,g.gentile,mirizzi,d.romito}@deemail.poliba.it,
{t.dinoia,disciascio}@poliba.it

Abstract. The recent spread of the so called Web of Data has made available a vast amount of interconnected data, paving the way to a new generation of ubiquitous applications able to exploit the information encoded in it. In this paper we present Cinemappy, a location-based application that computes contextual movie recommendations. Cinemappy refines the recommendation results of a content-based recommender system by exploiting contextual information related to the current spatial and temporal position of the user. The content-based engine leverages graph information within DBpedia, one of the best-known datasets publicly available in the Linked Open Data (LOD) project.

Keywords: Context-aware Recommender Systems, DBpedia, Linked Data, Movie Recommendation.

1 Introduction

Context can be defined as "*any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves*" [10]. This concept is particularly relevant whenever a user needs to look for information that does not depend exclusively from the particular knowledge domain. Thanks to great technological advances occurred in the latest years, particularly in ubiquitous computing, users are able to run almost any kind of application and to perform almost any task on small mobile devices. Smartphones and tablets are becoming a primary platform for information access [23]. If we think of a recommender task in a mobile scenario (e.g., choosing a movie in one of the nearest movie theaters, planning a sightseeing, etc.), we see that most recommendations are requested by users while they are on their way. This causes a continuous change in the context that needs to be carefully addressed. Recommendations are much more useful and enjoyable for end users as they change with their current context [6].

Recommender systems (RS) are information filtering and decision support tools addressing the problem of information overload, providing product and service recommendations personalized for user's needs and preferences.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 400–415, 2013.

In this paper we present an implementation of a content-based **C**ontext-**A**ware **R**ecommender **S**ystem (CARS) that gets information needed to recommend items from `Linked Open Data` (`LOD`) datasets [14] and combines it with other information freely available on the Web. In particular, our system recommends movies to be watched in theaters that are located in the user's neighborhood. The graph-based recommendation leverages `DBpedia` [7] as the knowledge base whence we extract movie features, such as genres, actors, directors, subjects, etc.. Main contributions of this paper are: (a) a semantic context-aware recommender system. In our approach we use semantic repositories and the notion of context to provide users with meaningful recommendations in a mobile environment; (b) the exploitation of heterogeneous information sources for movie recommendation. In particular we leverage data coming both from the `Web of Data` and from the traditional Web; (c) `Cinemappy` [19]: a proof-of-concept Android application exposing geo-location and context-aware features for movie recommendations in user neighborhood.

The remainder of this paper is organized as follows. In Section 2 we give some background information about the notion of context in recommender systems and how we exploit it in our approach. In Section 3 we present some basic notions about the usage of `DBpedia` as knowledge base for recommender systems. Then, in Section 4 we detail our approach and we present our mobile app. In Section 5 we discuss relevant related work. Conclusion and future work close the paper.

2 Context-Aware Recommender Systems in Mobility

A CARS deals with modeling and predicting user preferences by incorporating available contextual information into the recommendation process. Preferences and tastes are usually expressed as ratings and are modeled as a function of items, users and context. Hence, the rating function can be defined as:

$$r : User \times Item \times Context \rightarrow Rating$$

with the obvious meaning for *User* and *Item*. *Context* is defined by means of different attributes and criteria (we will detail them in the following) while *Rating* is usually represented as a Likert scale (going from 5 to 10 different values) or as a *Like/Don't like* Boolean set.

As in [20], we assume that there is a predefined finite set of contextual types in a given application and each of these types has a well-defined structure. In particular, in our mobile scenario we consider the context as represented by the following information:

Companion. There are many situations where a place or a service is more or less enjoyable depending on the people that are together with the user. Maybe the user and his/her friends love romantic movies but this is not the case of his/her partner. So, it would be fine if a movie recommender engine suggested romantic movies when the user is with his/her friends and comedies when he/she is with his/her partner.

Time. This is another important feature to consider. For example, in a movie theater recommender system, all the movies scheduled before the current time, plus the time to get to the theatre, have to be discarded.

Geographic Relevance. Geo-localized information plays a fundamental role in mobile applications. Depending on the current location of the user, a recommender engine should be able to suggest items close to them and discard the farther ones even if they may result more appealing with respect to the query. A location-aware recommender system should be able to suggest items or services for a given user whose location is known, considering more useful criteria than simply distance information. In [9] the authors propose ten criteria of geographic relevance. In the following we describe five of them that we considered relevant for our mobile application:

- ***Hierarchy:*** it represents the degree of separation between the current position of the user and that of the suggested item within a predefined spatial hierarchy. The main assumption is that geographic units are cognitively and empirically organized into a nested hierarchical form (e.g., city districts).

- ***Cluster:*** it is the degree of membership of an entity to a spatial cluster of related or unrelated entities. The user might be more interested in visiting a mall than a single shop.

- ***Co-location:*** usually users prefer locations where they may find other useful entities co-located with the one representing their main interest. As an example, it is common to have restaurants close to cinemas (since people like to go for dinner before watching or after having watched a movie).

- ***Association Rule:*** this criterion represents possible association rules that relates an entity with a related collection of geographic entities. The rules may comprise not only spatial information but also other kind of data (e.g., temporal) or their combination.

- ***Anchor-Point Proximity:*** this notion is related to the concept of landmarks. There are several key locations, such as our home and work place, that we consider as *"anchor"* points in our understanding of the geographic environment where we live. In general, we may define an anchor-point as a frequently visited location or a location where one spends a lot of time.

In order to enhance recommender systems results, context may be used in different ways. In [2], the authors identify three forms of context-aware recommendation systems: Contextual pre-filtering (*PreF*), Contextual post-filtering (*PoF*) and Contextual modelling. In Section 4 we describe in more detail the first two.

3 Feeding a Content-Based RS Using the Web of Data

In the recent years, thanks to the Web of Data advance, we are witnessing a flourishing of semantic datasets freely available on the Web encoding machine-understandable RDF[1] triples related to different domains and sometimes representing different points of view on the same domain. All this information can be exploited to model items and user profiles in an LOD-enabled content-based recommender system. One of the issues related to content-based approaches is the retrieval and pre-processing of the information used by the recommendation engine. In content-based (CB) recommender systems, the module in charge of extracting relevant information from items description and representing it as a vector of keywords, is the so called *Content Analyzer* (CA) [15]. It usually uses some Natural Language Processing techniques to extract/disambiguate/expand keywords in order to create a model of the item description. The use of LOD datasets to retrieve information related to an item eases the pre-processing steps performed by the CA since the information is already structured in an ontological way. Moreover, depending on the dataset, there is the availability of data related to diverse knowledge domains. If we consider datasets such as DBpedia or Freebase, we are able to access to rich linked data referring to a high variety of topics. Thanks to their SPARQL[2] endpoints, we can quite easily extract portions related to the movie domain from LOD datasets. We use this information as the base for our content-based recommender system.

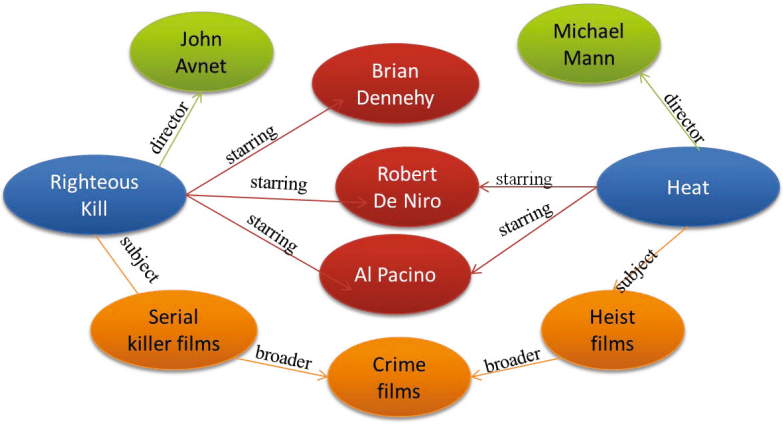


Fig. 1. A sample of an RDF graph related to the movie domain

[1] http://www.w3.org/RDF/
[2] http://www.w3.org/TR/rdf-sparql-query/

3.1 Computing Item Similarities in DBpedia

The main assumption behind our approach is that if two movies share some information (e.g., part of the cast, the director, the genre, some categories, etc.), then they are related with each other. Roughly speaking, the more features two movies have in common, the more they are similar. In a few words, a similarity between two movies (or two resources in general) can be detected if in the RDF graph: (1) they are directly related; (2) they are the subject of two RDF triples having the same property and the same object; (3) they are the object of two RDF triples having the same property and the same subject. Moreover, we exploit the ontological structure of the information conveyed within the Wikipedia categories, modeled in DBpedia by the properties dcterms:subject and skos:broader. This allows us to catch implicit relations and hidden information, i.e., information that is not directly detectable just looking at the nearest neighbors in the RDF graph.

Figure 1 shows a sample of the RDF graph containing properties and resources coming from DBpedia. In order to compute the similarities between movies, we adapted to an LOD-based setting one of the most popular models in classic information retrieval: the Vector Space Model (VSM). In this way we are able to represent items (i.e., movies) by means of feature vectors and then we can compute the similarity between them. In Section 4.1 we will provide more details on the use of DBpedia by the recommendation engine.

4 Cinemappy: A Context-aware Content-Based RS

In this section we describe Cinemappy, a mobile application that implements a context-aware recommender engine. The purpose is to suggest movies and movie theaters to the users based on their profile and on their current location (both spatial and temporal). On the one side, the context-aware section of the system is implemented by adopting both a *PreF* and a *PoF* approach. In order to retrieve all the data needed to evaluate the geographical criteria presented in Section 2, the application leverages information from other freely available Web sources such as *Google Places*[3] or *Trovacinema*[4]. On the other side, the CB part of the recommendation engine exploits the DBpedia graph structure for computing item similarities as described in Section 3. Moreover, driven by the context, the system also selects the right localized graph in DBpedia. Indeed, DBpedia contains also information extracted from localized versions of Wikipedia. Data coming from these Web sources are represented as different RDF graphs that can be easily selected via the FROM clause of a SPARQL query. The localized versions of DBpedia are particularly useful for the purpose of Cinemappy since some movies have for example a page in the Italian version of Wikipedia but they do not have a corresponding article in the English version. The basic building blocks of the system are depicted in Figure 2. All the data about the descriptions of

[3] http://www.google.com/places/

[4] This is an Italian Web site where you can find information related to cinemas and scheduled movies – http://trovacinema.repubblica.it/

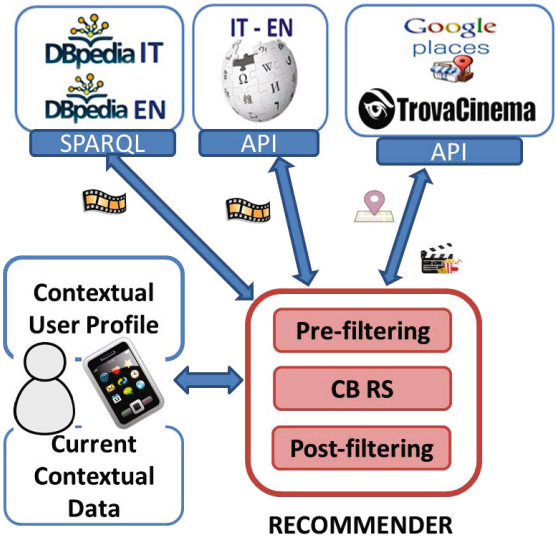


Fig. 2. System architecture

the movies are extracted via the `DBpedia SPARQL` endpoint. In particular, in our current implementation, we use both the *EN* and the *IT* graph since the application has been designed to be used by both English and Italian users. Information about theaters and movies is extracted from *Trovacinema* while geographic data (latitude/longitude) about theaters has been obtained using the *Google Geocoding API*[5]. In accordance with the *Co-location* principle, the system suggests POIs (Point of Interests) that are close to a given theater. In fact, the user might be interested in places close to the cinema such as restaurants, bars or playgrounds (if they are with their kids). The lists of POIs leverage the *Place Search* service that is part of the *Google Places API*[6]. These lists are created considering the relative distances between theaters and POIs and they use the *Google Distance Matrix API*[7]. In particular `Cinemappy` shows the POIs that are in a range of 2 km (about 1.24 miles) from a selected theater.

4.1 The Recommender Engine

`Cinemappy` uses several recommendation approaches to suggest movies and theaters to the user. Concerning contextual information we leverage both pre-filtering and post-filtering techniques for the contextual attributes introduced in Section 2. In particular to model the *Companion* attribute we use the so called

[5] `https://developers.google.com/maps/documentation/geocoding/`
[6] `https://developers.google.com/places/documentation/`
[7] `https://developers.google.com/maps/documentation/distancematrix/`

micro-profiling approach [4], a particular pre-filtering technique. Basically, with *micro-profiling* we associate a different profile to the user depending on the selected companion. Also *Time* is used to pre-filter recommendation results. For geographical data we use a post-filtering approach. Between pre-filtering and post-filtering phases, to match movies with contextual user-profiles, we use the content-based recommendation strategy which leverages `DBpedia` as the only information source. Before we continue in the description of the system, a few words need to be spent on how we model the user profile. In our setting, the user profile is based on a binary rating such as *Like/Don't like* (as the one adopted by YouTube). Empirical studies on real uses cases, as the one reported in the official YouTube Blog[8], show that even if users are allowed to rate an item on a five stars scale, usually it happens that they either assign the highest score or do not give any feedback at all. Therefore, we model ratings using a binary scale.

Contextual Pre-filtering. With `Cinemappy` we recommend bundles of items: *movies* to watch in *cinemas*. For this reason, movies that will not be featured in the future will not be suggested to the user. Nevertheless, such movies will be considered in the user profile if the user rated them. Moreover, for the current temporal and spatial position of the user, we constrain the set of movies to recommend considering geographical and time criteria. For each user u, the set of movies M_u is defined as containing the movies scheduled in the next d days in theaters in a range of k kilometers around the user position. The final recommendation list for u will be computed by considering only items available in M_u. This kind of restriction on the items with respect to time is a pre-filtering of the item set and not of the ratings as it usually happens in pre-filtering approaches. Regarding the companion context, the micro-profiling approach is modeled by considering a specific profile for u for each companion *cmp*:

$$profile(u, cmp) = \{\langle m_j, v_j \rangle \mid v_j = 1 \text{ if } u \text{ likes } m_j \text{ with companion } cmp,\ v_j = -1 \text{ otherwise}\}$$

In this way we are able to apply straightly the pre-filtering approach. When the user needs recommendations, given their current companion, the service considers only the corresponding micro-profile.

Content-Based Recommender. The recommendation algorithm is based on the one proposed in [12], enhanced with micro-profiles management. For the sake of completeness we briefly report here the main elements of the approach. In order to compute the similarities between movies, the Vector Space Model (VSM) [24] is adapted to an `LOD`-based setting. In VSM non-binary weights are assigned to index terms in queries and in documents (represented as sets of terms), and are used to compute the degree of similarity between each document in a collection and the query. In [12] the VSM, usually used for text-based retrieval, is

[8] `http://youtube-global.blogspot.it/2009/09/five-stars-dominate-ratings.html`

adapted in order to to deal with RDF graphs. In a nutshell, the whole RDF graph is represented as a 3-dimensional matrix where each slice refers to an ontology property and represents its adjacency matrix. A component (i.e. a cell in the matrix) is not *null* if there is a property that relates a subject (on the rows) to an object (on the columns). Given a property, each movie is seen as a vector, whose components refer to the *term frequency-inverse document frequency* TF-IDF (or better, in this case, *resource frequency-inverse movie frequency*). For a given slice (i.e. a particular property), the similarity degree between two movies is the correlation between the two vectors, and it is quantified by the cosine of the angle between them. All the nodes of the graph are represented both on the rows and on the columns of the matrix. A few words need to be spent for the properties `dcterms:subject` and `skos:broader` which are very popular, e.g., in the `DBpedia` dataset. As also shown in Figure 1, every movie is related to a category by the property `dcterms:subject` which is in turn related to other categories via `skos:broader` organized in a hierarchical structure. To the purpose of the recommendation, `skos:broader` is considered as *one-step transitive*. We will explain this notion with the aid of a simple example. Suppose to have the following RDF statements:

```
dbpedia:Righteous_Kill
    dcterms:subject dbpedia:Category:Serial_killer_films .
dbpedia:Category:Serial_killer_films
    skos:broader dbpedia:Category:Crime_films .
```

Starting from `dbpedia:Category:Serial_killer_films` we have that `dbpedia:Category:Crime_films` is at a distance of one step. Hence, by considering a *one-step transitivity*, we have that:

```
dbpedia:Righteous_Kill
    dcterms:subject dbpedia:Category:Crime_films .
```

is inferred by the original statements.

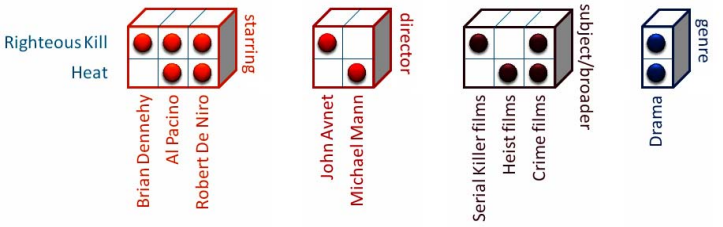


Fig. 3. Matrix representation of the RDF graph of Figure 1

By looking at the model, we may say that: (1) the matrix is very sparse; (2) properties are considered as independent with each other (there is no `rdfs:subPropertyOf` relation); (3) the focus is on discovering the similarities

between movies (or in general between resources of the same `rdf:type` and not between each pair of resources). Based on the above observations, the matrix slices can be decomposed into smaller matrices where each matrix refers to a specific `RDF` property, as shown in Figure 3. In other words, for each matrix, the rows represent somehow the *domain* of the considered property, while the columns its *range*. For a given property, the components of each row represent the contribution of a resource (i.e. an actor, a director, etc.) to the corresponding movie. With respect to a selected property p, a movie m is then represented by a vector containing all the terms/nodes related to m via p. As for classical Information Retrieval, the index terms $k_{n,p}$, that is all the nodes n linked to a movie by a specific property p, are assumed to be all mutually independent and are represented as unit vectors of a t-dimensional space, where t is the total number of index terms. Referring to Figure 3, the index terms for the *starring* property are *Brian Dennehy*, *Al Pacino* and *Robert De Niro*, while $t = 3$ is the number of all the actors that are objects of a triple involving *starring*. The representation of a movie m_i, according to the property p, is a t-dimensional vector given by $\overrightarrow{m_{i,p}} = (w_{1,i,p}, w_{2,i,p}, ..., w_{t,i,p})$, where $w_{n,i,p}$ is a non-negative and non-binary value representing the weight associated with a term-movie pair $(k_{n,p}, \overrightarrow{m_{i,p}})$. The weights $w_{n,i,p}$ adopted in the model are TF-IDF weights. More precisely, the TF ($f_{n,i,p}$) is the frequency of the node n, as the object of an `RDF` triple having p as property and the node i as subject (the movie). Actually, this term can be either 0 (if i is not related to n via p) or 1, since two identical triples can not coexist in an `RDF` graph. As for classical information retrieval, the IDF is computed as the logarithm of the ratio between M, that is the total number of movies in the collection, and $a_{n,p}$, that is the number of movies that are linked to the resource n, by means of the predicate p. As an example, referring to Figure 3, for the *starring* property, and considering $n = AlPacino$, then $a_{AlPacino,starring}$ is equal to 2, and it represents the number of movies where *Al Pacino* acted. Relying on the model presented above, each movie can be represented as a $t \times P$ matrix, where P is the total number of selected properties. If we consider a projection on a property p, each pair of movies, m_i and m_j, are represented as t-dimensional vectors. The degree of similarity between m_i and m_j with respect to p is evaluated as the correlation between the vectors $\overrightarrow{m_{i,p}}$ and $\overrightarrow{m_{j,p}}$. More precisely, the correlation is computed as the cosine of the angle between the two vectors:

$$sim^p(m_i, m_j) = \frac{\sum_{n=1}^{t} w_{n,i,p} \cdot w_{n,j,p}}{\sqrt{\sum_{n=1}^{t} w_{n,i,p}^2} \cdot \sqrt{\sum_{n=1}^{t} w_{n,j,p}^2}}$$

The method described so far is general enough and it can be applied when the similarity has to be found between resources that appear as subjects of `RDF` triples. When the resources to be ranked appear as objects of `RDF` triples, it is simply a matter of swapping the rows with the columns in the matrices of Figure 3 and applying again the same algorithm. Lastly, when two resources are directly related by some specific properties (as the case of the property `dbpedia:subsequentWork`), it is sufficient to operate a matrix transformation

to handle this case in the same way as done so far. In the following we will see how to combine such similarity values with a user profile to compute a content-based recommendation. In order to evaluate if a movie $m_i \in M_u$ might be of interest for u given *cmp* we need to combine the similarity values related to each single property p of m_i and compute an overall similarity value $\tilde{r}_{PreF}(u_{cmp}, m_i)$:

$$\tilde{r}_{PreF}(u_{cmp}, m_i) = \frac{\sum_{m_j \in profile(u,cmp)} v_j \times \frac{\sum_p \alpha_p \times sim^p(m_j, m_i)}{P}}{|profile(u, cmp)|}$$

where P represents the number of properties in `DBpedia` we consider relevant for our domain (e.g. `dbpedia-owl:starring`, `dcterms:subject`, `skos:broader`, `dbpedia-owl:director`) and $|profile(u, cmp)|$ is the cardinality of the set $profile(u, cmp)$. A weight α_p is assigned to each property representing its worth with respect to the user profile. In order to compute these weights, supervised machine learning techniques are adopted. In particular we use a genetic algorithm to find the optimal weights. We train our model on `MovieLens`[9], a popular dataset in the movie domain.

Based on $\tilde{r}_{PreF}(u_{cmp}, m_i)$, we compute the ranked list $R_{u_{cmp}}$ of potential movies that will be suggested to the user.

Contextual Post-filtering. Based on geographical criteria, we apply post-filtering on $R_{u_{cmp}}$ to re-rank its elements. In particular, for each criterion we introduce a $\{0, 1\}$-variable whose value is defined as follows:
`h` (*hierarchy*) : it is equal to 1 if the cinema is in the same city of the current user position, 0 otherwise;
`c` (*cluster*): it is equal to 1 if the cinema is part of a multiplex cinema, 0 otherwise;
`cl` (*co-location*): it is equal to 1 if the cinema is close to other POIs, 0 otherwise;
`ar` (*association-rule*): it is equal to 1 if the user knows the price of the ticket, 0 otherwise. This information is caught implicitly from the information about the cinema;
`ap` (*anchor-point proximity*): it is equal to 1 if the cinema is close to the user's house or the user's office, 0 otherwise.
These geographic criteria are combined with $\tilde{r}_{PreF}(u_{cmp}, m_i)$ to obtain a single score:

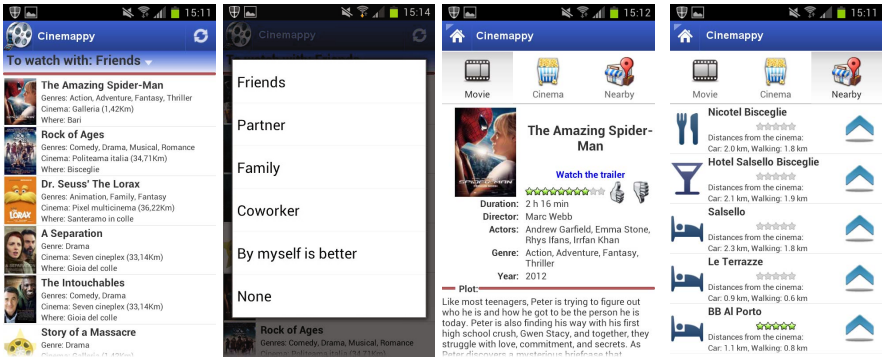
$$\tilde{r}(u_{cmp}, m_i) = \beta_1 \times \tilde{r}_{PreF}(u_{cmp}, m_i) + \beta_2 \times \frac{(\texttt{h} + \texttt{c} + \texttt{cl} + \texttt{ar} + \texttt{ap})}{5}$$

where $\beta_1 + \beta_2 = 1$. In the current implementation of `Cinemappy`, both β_1 and β_2 have been chosen experimentally and have been set respectively to 0.7 and 0.3.

[9] `http://www.grouplens.org/node/12`

4.2 Implementation

Cinemappy has been implemented as a mobile application for Android smartphones[10]. the user starts the application, Cinemappy displays a list of movies



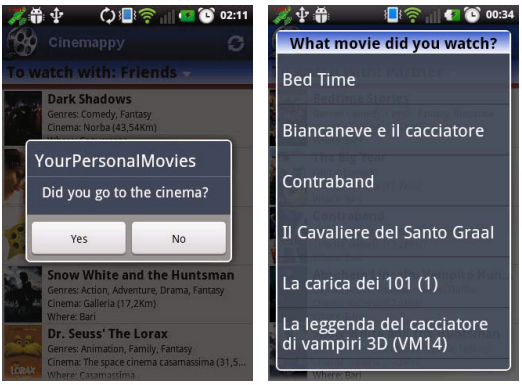
(a) (b) (c) (d)

Fig. 4. Some screenshots of Cinemappy

according to the current contextual user profile (Figure 4(a)). The user can choose their current companion from a list of different options thus enabling their micro-profile (Figure 4(b)). Please note that even if the different user micro-profiles are empty, Cinemappy is able to suggest movies based exclusively on contextual information. For each movie in the list, its genres and the distance of the suggested theater from the user position are shown. Hence, the user can click on one of the suggested movies and look at its description, watch its trailer and express a preference in terms of *I would watch*/*I would not watch* (Figure 4(c)). Furthermore, the user can find information about the recommended theater or the other theaters that feature that movie. Based on the theater location, the user could be interested in places where spending time with their friends, such as pubs, or with their girlfriend/boyfriend such as restaurants or bars, or with their family, and in this case maybe the user could be interested in certain kind of places also adequate for children. To support the user in this choice, the application suggests POIs by considering contextual criteria (Figure 4(d)).

Location-Based Rating Acquisition Service. User preferences acquisition plays a very important role for recommender systems in real scenarios. As previously pointed out, the system allows the user to rate movies while they are looking at their descriptions. Furthermore, thanks to the ubiquitous-awareness of mobile systems we are able to ask users to elicit their preferences in a more pervasive way. We exploit the geo-localization capabilities of mobile devices to

[10] https://play.google.com/store/apps/details?id=it.sisinflab.lod.mobile.cinemappy&hl=en_GB



(a) (b)

Fig. 5. Location-Based Rating Acquisition Service

understand if the user watched a movie. Every 90 minutes the application, by means of a background service, captures the user position. If the user has been for at least 90 minutes in a similar position close to a cinema in a time span corresponding to one or more scheduled movies, we assume that the user watched a movie in that cinema. In this case, the application asks the user if they went to the cinema (Figure 5(a)) and and if a positive answer ensues, the user can rate one of the movies featured in that cinema (Figure 5(b)).

5 Related Work

In this section we report on some of the main related works that we consider relevant to our approach and we classify them as *Context-Aware and Mobile Recommender Systems* and *Semantics in Recommender Systems*. A complete literature review in the two fields is out of the scope of this paper.

5.1 Context-Aware and Mobile Recommender Systems

As argued in [1], incorporating contextual information in traditional RSs is very important if we want to increase the quality of returned results. In the paper, the authors describe a multidimensional approach to recommendations, wherein the traditional user-item matrix is extended to a multidimensional model by adding new elements such as *place*, *time*, etc.. The context-aware recommendation process can take one of the three forms, depending on which stage of the process the context is applied in [2]. These forms are: *Contextual pre-filtering*, *Contextual post-filtering* and *Contextual modeling*. The *post-filtering* and *post-filtering* methods are compared in [21] where the authors, based on some experimental results, propose a simple but effective and practical way to decide how to

use the two methods in a recommendation engine. Context-aware recommender systems have attracted a lot of attention in mobile application. Mobile phones allow users to have access to a vast quantity of information in an ubiquitous way. In [23], the author details issues, scenarios and opportunities regarding mobile RSs especially in the area of travel and tourism. He describes the major techniques and specific computational models that have been proposed for mobile recommender systems. In [25] the authors describe *COMPASS*, a context-aware mobile tourist application where the context is modelled as the current user's requests. With regards to the profile, needs and context information of the user, *COMPASS* performs a selection of potentially interesting nearby buildings, buddies and other objects. These information change whenever the user moves or they change their goal. In [5] the authors present *ReRex*, a context-aware mobile recommender system that suggests POIs. By means of a Web-based survey application, the users are requested to imagine a contextual condition and then to evaluate a POI. In the answer the users have in mind the effect of the context on their decisions. The authors built a predictive model that is able to predict the relevance of such a POI in the contextual condition. Then, the application uses this model to allow the user to select the contextual factors and to browse the related context-aware recommendations.

5.2 Semantics in Recommender Systems

The need for a semantic representation of data and user profiles has been identified as one of the next challenges in the field of recommender systems [16]. Some ontology-based RSs are presented in [16] where the authors describe an approach that exploits the usage of ontologies to compute recommendations. Two systems, *Quickstep* and *Foxtrot*, are introduced that make use of semantic user profiles to compute collaborative recommendations. The profiles are represented by topics about research papers with respect to an ontology and the recommendations are computed matching the topics of the current user profile with the topics of similar users' profiles. The authors prove that: ontological inference improves the user profiling; the ontological knowledge facilitates the initial bootstrap of the system resolving the cold-start problem; the profile visualization improves the user profiling accuracy. A hybrid recommendation system is proposed in [8] wherein user preferences and item features are described by semantic concepts. These latter are clustered in order to obtain user's clusters corresponding to implicit Communities of Interest. In [18] the authors introduce the so called *semantically enhanced collaborative filtering* in which structured semantic knowledge about items is used in conjunction with user-item ratings to create a combined similarity measure for item comparisons. In [3] an approach is presented that integrates user rating vectors with an item ontology. In these works, the experiments prove an accuracy improvement over classical collaborative approaches also on the presence of sparse datasets. Most of the works described so far have been produced when `LOD` did not exist. The Web Of Data

paves the way to the usage of new and rich semantic datasets to compute recommendations. In [13] the authors present a generic knowledge-based description framework built upon semantic networks. The aim of the framework is to integrate and to exploit some knowledge on several domains in order to compute cross-domain recommendations. They use a spreading activation method with the purpose of finding semantic relatedness between items belonging to different domains. *dbrec* [22] is a music content-based recommender system leveraging the `DBpedia` dataset. They define the *Linked Data Semantic Distance* in order to find semantic distances between resources and then compute recommendations. In [11,12] a model-based approach and a memory-based one to compute CB recommendations are presented leveraging `LOD` datasets. Several strategies are described and compared to select ontological properties (in the movie domain) to be used during the computation of recommended items. A different hybrid technique is adopted in [17], where `DBpedia` is exploited as background knowledge for semantic tags recommendation. The semantic data coming from `DBpedia` are mixed with keyword-based information extracted via Web search engines to compute the semantic relatedness between the query posed by the user and the resources available in a semantic graph.

6 Conclusion and Future Work

In this work, we presented `Cinemappy`: a context-aware content-based recommender system for movies and movie theaters suggestions. The content-based part of the recommender engine is fed with data coming from localized `DBpedia` graphs and the results are enhanced by exploiting contextual information about the user. The application has been implemented as an Android application. Geographic criteria that go beyond the simple geographic distance have been implemented to fully exploit location-based information. Our future plans are: (a) evaluate the overall approach with real users; (b) enrich the information used by the content-based RS with other datasets in the `LOD` cloud; (c) apply the same approach to different context-aware domains such as tourist spots recommendations.

Acknowledgments. The authors acknowledge support of PON01_00850 ASK-HEALTH project.

References

1. Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating contextual information in recommender systems using a multidimensional approach. ACM Trans. Inf. Syst. 23(1), 103–145 (2005)
2. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Recommender Systems Handbook, pp. 217–253 (2011)
3. Anand, S.S., Kearney, P., Shapcott, M.: Generating semantically enriched user profiles for web personalization. ACM Trans. Internet Technol. 7(4) (October 2007)

4. Baltrunas, L., Amatriain, X.: Towards Time-Dependant Recommendation based on Implicit Feedback. In: Context-aware Recommender Systems Workshop at Recsys 2009 (2009)
5. Baltrunas, L., Ludwig, B., Peer, S., Ricci, F.: Context-aware places of interest recommendations for mobile users. In: Marcus, A. (ed.) HCII 2011 and DUXU 2011, Part I. LNCS, vol. 6769, pp. 531–540. Springer, Heidelberg (2011)
6. Baltrunas, L., Ludwig, B., Peer, S., Ricci, F.: Context relevance assessment and exploitation in mobile recommender systems. Personal and Ubiquitous Computing 16(5), 507–526 (2012)
7. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. Web Semant. 7, 154–165 (2009)
8. Cantador, I., Bellogín, A., Castells, P.: A multilayer ontology-based hybrid recommendation model. AI Commun. 21(2-3), 203–210 (2008)
9. De Sabbata, S., Reichenbacher, T.: Criteria of Geographic Relevance: An Experimental Study. International Journal of Geographical Information Science (2012)
10. Dey, A.K.: Understanding and using context. Personal Ubiquitous Comput. 5(1), 4–7 (2001)
11. Di Noia, T., Mirizzi, R., Ostuni, V.C., Romito, D.: Exploiting the web of data in model-based recommender systems. In: 6th ACM Conference on Recommender Systems (RecSys 2012). ACM, ACM Press (2012)
12. Di Noia, T., Mirizzi, R., Ostuni, V.C., Romito, D., Zanker, M.: Linked open data to support content-based recommender systems. In: 8th International Conference on Semantic Systems (I-SEMANTICS 2012). ICP. ACM Press (2012)
13. Fernández-Tobías, I., Cantador, I., Kaminskas, M., Ricci, F.: A generic semantic-based framework for cross-domain recommendation. In: Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems, HetRec 2011, pp. 25–32. ACM, New York (2011)
14. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers (2011)
15. Lops, P., Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. In: Recommender Systems Handbook, pp. 73–105 (2011)
16. Middleton, S.E., Roure, D.D., Shadbolt, N.R.: Ontology-based recommender systems. Handbook on Ontologies 32(6), 779–796 (2009)
17. Mirizzi, R., Ragone, A., Di Noia, T., Di Sciascio, E.: Ranking the linked data: the case of dbpedia. In: Benatallah, B., Casati, F., Kappel, G., Rossi, G. (eds.) ICWE 2010. LNCS, vol. 6189, pp. 337–354. Springer, Heidelberg (2010)
18. Mobasher, B., Jin, X., Zhou, Y.: Semantically enhanced collaborative filtering on the web. In: Berendt, B., Hotho, A., Mladenič, D., van Someren, M., Spiliopoulou, M., Stumme, G. (eds.) EWMF 2003. LNCS (LNAI), vol. 3209, pp. 57–76. Springer, Heidelberg (2004)
19. Ostuni, V.C., Di Noia, T., Mirizzi, R., Romito, D., Di Sciascio, E.: Cinemappy: a context-aware mobile app for movie recommendations boosted by dbpedia. In: 1st International Workshop on Semantic Technologies Meet Recommender Systems & Big Data (SeRSy 2012), vol. 919. CEUR-WS (2012)
20. Palmisano, C., Tuzhilin, A., Gorgoglione, M.: Using context to improve predictive modeling of customers in personalization applications. IEEE Trans. Knowl. Data Eng. 20(11), 1535–1549 (2008)

21. Panniello, U., Tuzhilin, A., Gorgoglione, M., Palmisano, C., Pedone, A.: Experimental comparison of pre- vs. post-filtering approaches in context-aware recommender systems. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys 2009, pp. 265–268. ACM, New York (2009)
22. Passant, A.: Measuring semantic distance on linking data and using it for resources recommendations. In: Proceedings of the AAAI Spring Symposium "Linked Data Meets Artificial Intelligence" (March 2010)
23. Ricci, F.: Mobile recommender systems. J. of IT & Tourism 12(3), 205–231 (2011)
24. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18, 613–620 (1975)
25. van Setten, M., Pokraev, S., Koolwaaij, J.: Context-aware recommendations in the mobile tourist application compass. In: De Bra, P.M.E., Nejdl, W. (eds.) AH 2004. LNCS, vol. 3137, pp. 235–244. Springer, Heidelberg (2004)

ChEA2: Gene-Set Libraries from ChIP-X Experiments to Decode the Transcription Regulome

Yan Kou, Edward Y. Chen, Neil R. Clark, Qiaonan Duan, Christopher M. Tan, and Avi Ma'ayan*

Department of Pharmacology and Systems Therapeutics,
Systems Biology Center New York (SBCNY), Icahn School of Medicine at Mount Sinai,
New York, NY 10029
avi.maayan@mssm.edu

Abstract. ChIP-seq experiments provide a plethora of data regarding transcription regulation in mammalian cells. Integrating ChIP-seq studies into a computable resource is potentially useful for further knowledge extraction from such data. We continually collect and expand a database where we convert results from ChIP-seq experiments into gene-set libraries. The manual portion of this database currently contains 200 transcription factors from 221 publications for a total of 458,471 transcription-factor/target interactions. In addition, we automatically compiled data from the ENCODE project which includes 920 experiments applied to 44 cell-lines profiling 160 transcription factors for a total of ~1.4 million transcription-factor/target-gene interactions. Moreover, we processed data from the NIH Epigenomics Roadmap project for 27 different types of histone marks in 64 different human cell-lines. All together the data was processed into three simple gene-set libraries where the set label is either a mammalian transcription factor or a histone modification mark in a particular cell line, organism and experiment. Such gene-set libraries are useful for elucidating the experimentally determined transcriptional networks regulating lists of genes of interest using gene-set enrichment analyses. Furthermore, from these three gene-set libraries, we constructed regulatory networks of transcription factors and histone modifications to identify groups of regulators that work together. For example, we found that the Polycomb Repressive Complex 2 (PRC2) is involved with three distinct clusters each interacting with different sets of transcription factors. Notably, the combined dataset is made into web-based application software where users can perform enrichment analyses or download the data in various formats. The open source ChEA2 web-based software and datasets are available freely online at http://amp.pharm.mssm.edu/ChEA2.

Keywords: ChIP-seq, ChIP-chip, Microarrays, Systems Biology, ENCODE, Enrichment Analysis, Transcriptional Networks, Data Integration, Data Visualization, JavaScript D3.

* Corresponding author.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 416–430, 2013.

1 Introduction

Gene expression in mammalian cells is regulated by transcriptional complexes that include transcription factors, histone modifiers, other chromatin and DNA modifiers, and co-regulators proteins that bring these factors together in a transient manner. These cellular components are increasingly profiled using the ChIP-seq/chip (ChIP-X) technologies to identify the locations on the genome where such cellular components bind. Integrating datasets from many sparse ChIP-X studies applied to mammalian cells into a single computable resource is challenging but can lead to a new level of global understanding of gene expression regulation in mammalian cells. In 2010, a database called ChIP-X Enrichment Analysis (ChEA) was published [1]. This database originally contained integrated data from 87 publications reporting the binding of 92 mammalian transcription factors to their putative targets. Since then, we continually expanded this database where it currently contains 200 transcription factors from 221 publications for a total of 458,471 transcription-factor/target-gene interactions. Such dataset is potentially useful for two types of applications: 1) exploring the global transcriptional regulatory landscape of mammalian cells; 2) identifying the most likely regulators given lists of differentially expressed genes using gene-list enrichment analyses.

While the initial version of the ChEA database has been proven useful for several applications, since its publication in 2010 other similar databases have been published. For example hmChIP provides systematic access to ChIP-X experiments deposited to GEO [2]. ChIP-Array combines gene expression changes together with ChIP-X experiments to provide querying capabilities [3]. Similarly to ChIP-Array, TranscriptomeBrowser integrates various types of binary regulatory interactions including those extracted from ChIP-X experiments [4]. Two other related tools that aggregate ChIP-X experiments are called CistromeMap [5] and CistromeFinder [6]. CistromeMap and CistromeFinder combine many ChIP-seq and DNase-seq publications for integrative analysis. Complimenting these efforts are databases and tools that identify transcription-factor/target-genes interactions by other computational and experimental methods. For example a database called HTRIdb aggregates mammalian transcription-factor/target-gene interactions from low-throughput studies [7] and there are many tools that detect such interactions computationally using position weight matrices. However, not much meta-analysis has been performed on the networks that such datasets and tools produce for the purpose of obtaining a global view of the transcriptional regulatory landscape in mammalian cells. In addition, the data from these tools is not easily utilized for gene-list enrichment analyses. Here we performed an initial integrative analysis of the ChIP-X data we collected, as well as provide a state-of-the-art HTML5 freely available web application that can be used to query the datasets with lists of differentially expressed genes by performing gene-list enrichment analyses.

2 Methods

2.1 The Histone Modification Dataset

ChIP-seq for histone modification datasets were collected from the Roadmap Epigenomics project (http://www.roadmapepigenomics.org) deposited to the GEO

database (http://www.ncbi.nlm.nih.gov/geo/). Previous studies [8] have indicated that the use of control samples substantially reduces DNA shearing biases and sequencing artifacts, therefore for each experiment, an input control sample was matched according to the description in GEO. ChIP-seq experiments without matched control dataset were not included. BED files were first downloaded and standardized to contain the first six columns as defined by the UCSC genome browser BED file format by adding placeholders if needed. Unlike narrow peaks of transcription factor binding sites, histone modification marks have broad peaks and can be extended to as long as 100kb [8]; therefore the peak-calling software Sicer [9] was used to identify significant peaks. Sicer uses a spatial clustering approach that considers enrichment information from neighboring regions to call broader domains. Each read was shifted by 75bp towards the center of the DNA fragment; window size of 200bp and gap size equals to the window size, or three times the window size, were used according to the authors' guide. However, for large datasets, window size of 500bp and gap size of 1500bp were applied because smaller parameters resulted in memory limit errors. For broad histone marks such as H3K27me3 and H3K36me3, the window size of 500bp and gap size of 1500bp were set to detect significant peaks. Peaks were identified with only one read mapped to the reference genome, not allowing for redundant reads. False discovery rate (FDR) of 0.001 was used to find significant peaks as compared to the control.

The significant peaks for each experiment were collected and associated with annotated human genes. Chromosome coordinates of the human hg19 build were downloaded from NCBI; the longest readout was kept for genes having multiple isoforms. A gene is considered associated with a significant peak if the peak overlaps with the region from -10kb of transcription starting site (TSS) to +5kb of transcription ending site (TES) of the gene. Peaks that overlap with $\pm$2kb of TSS of known genes were defined as enriched at promoter regions. Genes that have significant peaks at promoter regions were ranked according to their distance from the middle point of the associated peak to the transcription starting site. For experiments that identified more than 2000 genes around TSS and TES, the top 2000 genes were taken to generate the gene matrix transpose (GMT) gene-set library file, whereas all genes were taken for experiments that recovered less genes. The resulted dataset contains 27 types of histone modifications for 64 human cell lines and tissue types.

2.2 The ENCODE Transcription Factor Dataset

ChIP-seq datasets for transcription factor binding sites were downloaded from the ENCODE project [10] (ftp://encodeftp.cse.ucsc.edu/pipeline/hg19/) on July 9, 2012. A total of 920 experiments applied to 44 cell lines profiling 160 transcription factors were processed. For each experiment, the peak file in broadPeak or narrowPeak format was extracted. For each experiment dataset, the peaks in the broadPeak and narrowPeak files were mapped to the TSS of human genes from the hg19 build if the peak overlaps with the region of (-5kb, +2kb) surrounding the TSS. Most of the experiments have biological replicates. Therefore, only the unique genes (intersection) with peaks deposited within the defined promoter region from both replicates were extracted for further analysis. The GMT file was generated in the same way as for the histone modification datasets, where the top 2000 genes were taken for experiments that yielded larger number of target genes.

2.3 Updating the Manual Part of the ChEA Database

The manual part of the ChEA database is continually updated by manually processing the supporting tables from ChIP-seq publications. Target gene lists are extracted from such publications. Only human or mouse experiments are considered. All gene identifiers are converted to Entrez Gene Symbols. E-mail alerts for abstracts that contain the key terms "ChIP-seq" or "ChIP-chip" link to publications that may contain data that is suitable for this database. In some cases when the authors only provide the peaks' coordinates we process the data to identify and rank the associated genes.

2.4 The ChEA2 Web Application

ChEA2 uses HTML5, CSS3, and JavaScript on the frontend to render the page, make the SVG visualizations interactive, and retrieve JSON data from the backend server asynchronously using AJAX. The frontend makes heavy use of the jQuery JavaScript library and other dependent plugins to simplify page manipulation/selection, enable touch gestures, add sortable tables, and create tooltips in combination with using the D3 JavaScript library to generate SVG visualizations. ChEA2 also utilizes JSP to generate some of the more dynamic pages that interact with the backend. The backend server side uses an Apache Tomcat 7 application server to host Java servlet web apps and a Microsoft IIS 6 web server to serve the web pages. Apache Maven is used to simplify the process of compiling the code, processing the web resources, and deploying the web app to the Tomcat server. The application is best viewed with Chrome or Firefox, but it is compatible with all modern browsers that support SVG and HTML5. The code-based is mostly borrowed from the recently published tool Enrichr [11].

2.5 Canvas Visualization

The canvas images in Figure 1 represent an alternative way to visualize networks. Instead of connecting nodes with edges, the network nodes are placed onto a square toroidal grid and then clustered together using simulated annealing to maximize local connections. The brightness of a node on the finished grid corresponds to the strength of its connections with its neighbors. The annealing process is performed by swapping two randomly selected nodes on the grid and recalculating the global fitness. After annealing, the canvas is exported as a JavaScript Object Notation (JSON) file which is then sent to the Visualizer module. The Visualizer creates SVG images with HTML5 and the JavaScript Library D3. These images were created with the recently published tool, Network2Canvas [12].

3 Results

The manual portion of the ChEA2 database currently contains 200 transcription factors from 221 publications for a total of 458,471 transcription-factor/target interactions. In addition, the data processed from ENCODE includes 920 experiments applied in 44 cell-lines profiling 160 transcription factors for a total of ~1.4 million transcription-factor/target-gene interactions. The data from the NIH Epigenomics

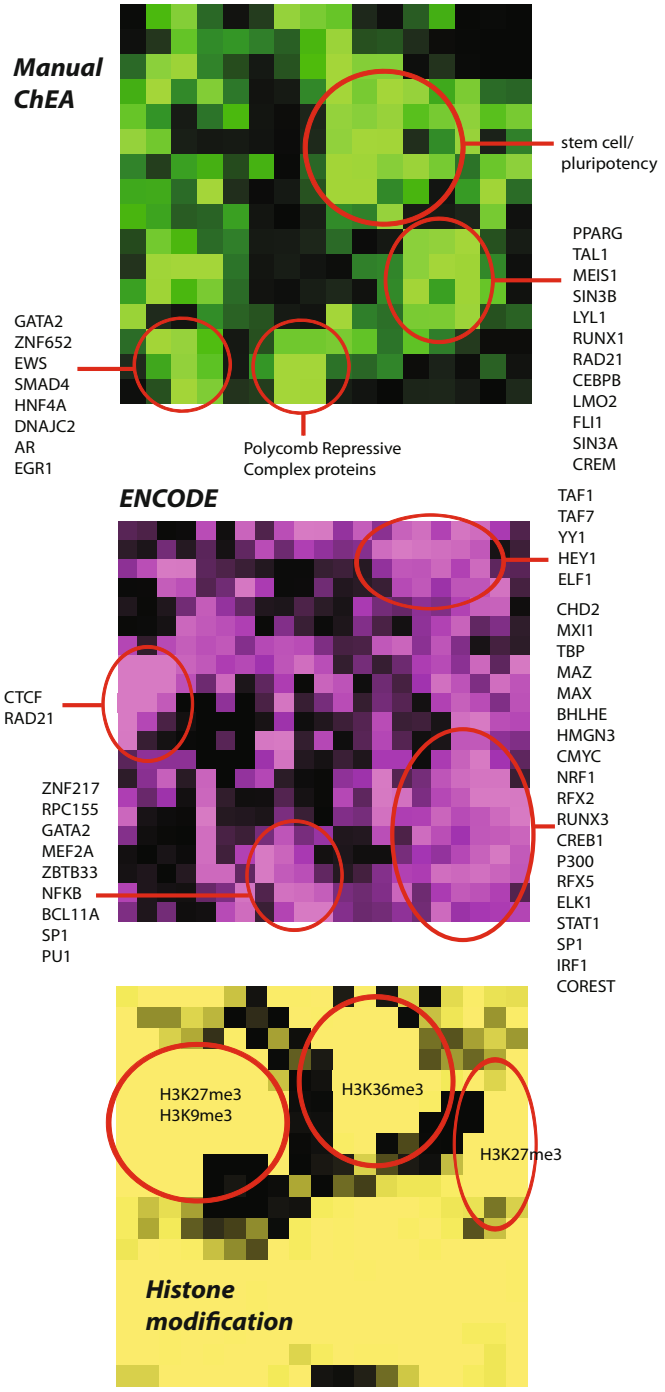


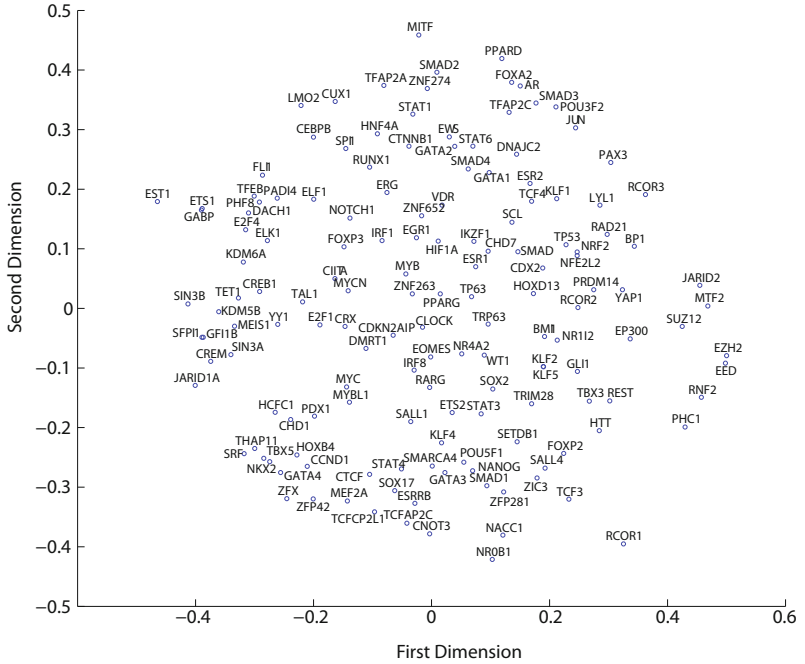
Fig. 1. Canvas representation of transcription factors and histone modifications networks using Network2Canvas (N2C) [12]

Roadmap project contain 27 different types of histone marks in 64 different human cell-lines for a total of 438,064 interactions. The first step in our analysis is to visualize each of the gene-set libraries: the manual ChEA, ENCODE, and histone modifications on a canvas. The canvas visualization is a compact alternative to visualize networks as ball-and-stick diagrams [12]. In these canvases, each ChIP-X experiment is represented as a square on a grid of squares. The location of the experiments/squares is optimized so a gene-set with a similar gene-set are adjacent to each other. The similarity measure for the three grids is based on set similarity computed by the Sets2Networks algorithm [13]. Once such arrangement is determined, the squares are color-coded by the level of similarity each experiment/square has with its neighbors. The brighter the square the more similar it is to its neighboring squares. The canvases clearly identify clusters within the gene-set libraries/networks (Fig. 1). These clusters include a stem-cell/pluripotency set of factors and the PRC2 group in the manual ChEA canvas; the H3K36me3, H3K27me3, mixture of H3K27me3/H3K9me3 clusters in the histone modifications canvas; and CTCF/RAD21, TAF1/TAF7/YY1/HEY1 clusters in the ENCODE canvas; but there other clusters. While some of these clusters are likely functional, where the factors physically interact to regulate the same sets of genes, it should be considered that some factors are more frequently profiled so there is some research bias in the formation of the various clusters on these canvases.

Next, we visualized the manual ChEA gene-set library and the ENCODE gene-set library using multi-dimensional scaling (MDS) with a distance measure similarity computed using the Sets2Networks algorithm [13] (Fig. 2). This visualization identifies the same clusters but has the advantage of showing the labels for most factors as well as provides a visualization of the distance similarity among all factors. Then, we visualized the manual ChEA gene-set library as a heatmap using hierarchical clustering (Fig. 3). The heatmap identifies three clusters that include the members of the PRC2 complex: EZH2, EED, SUZ12, and JARID2. Besides these core components of PRC2, the transcription factor MTF2, and the histone methyltransferase SETDB1 as well as RNF2 and PHC1 which are both members of the PRC1 complex, cluster with the canonical PRC2 components. Interestingly, the transcription factor p53 also clusters with these factors. However, these three clusters also include various other factors that are unique for each one of the three clusters. Using known protein-protein interactions from the literature [14], we see that many of these additional factors were previously identified to interact with members of the PRC2 complex, suggesting that these factors directly physically associate with the PRC2 complex members to regulate the same set of genes in different contexts (Fig. 3, right). This potentially identifies multiple roles for the PRC2 complex in different cell types. As we see below, when combining the manual ChEA dataset with the ENCODE and histone modification datasets the PRC2 complex members strongly associates with the H3K27me3 modification which is already well established.

Next, we visualized the manual ChEA and ENCODE networks as ball-and-stick diagrams. We only connected the transcription factors that were used as immuno-precipitation baits in ChIP-X experiments (Fig. 4). Therefore all nodes are transcription factors and the edges connect transcription factor profiled by ChIP-X to their direct binding targets which are other transcription factors. The manual ChEA network is made of 197 factors connected through 5344 edges whereas the ENCODE network is made of 150 factors connected through 2430 links with clustering

ChEA



ENCODE

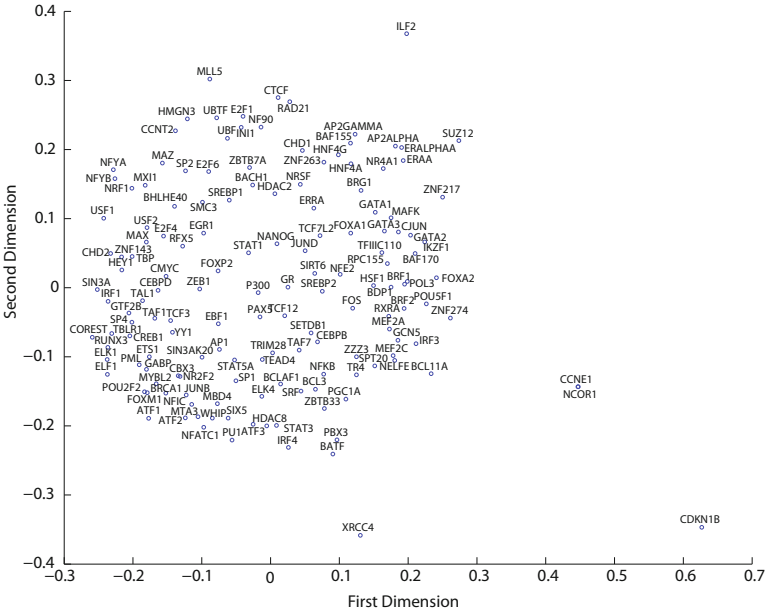


Fig. 2. Multi-dimensional scaling scatter plot of the a) manual ChEA; and b) ENCODE datasets, using the Sets2Networks distance measure [13]

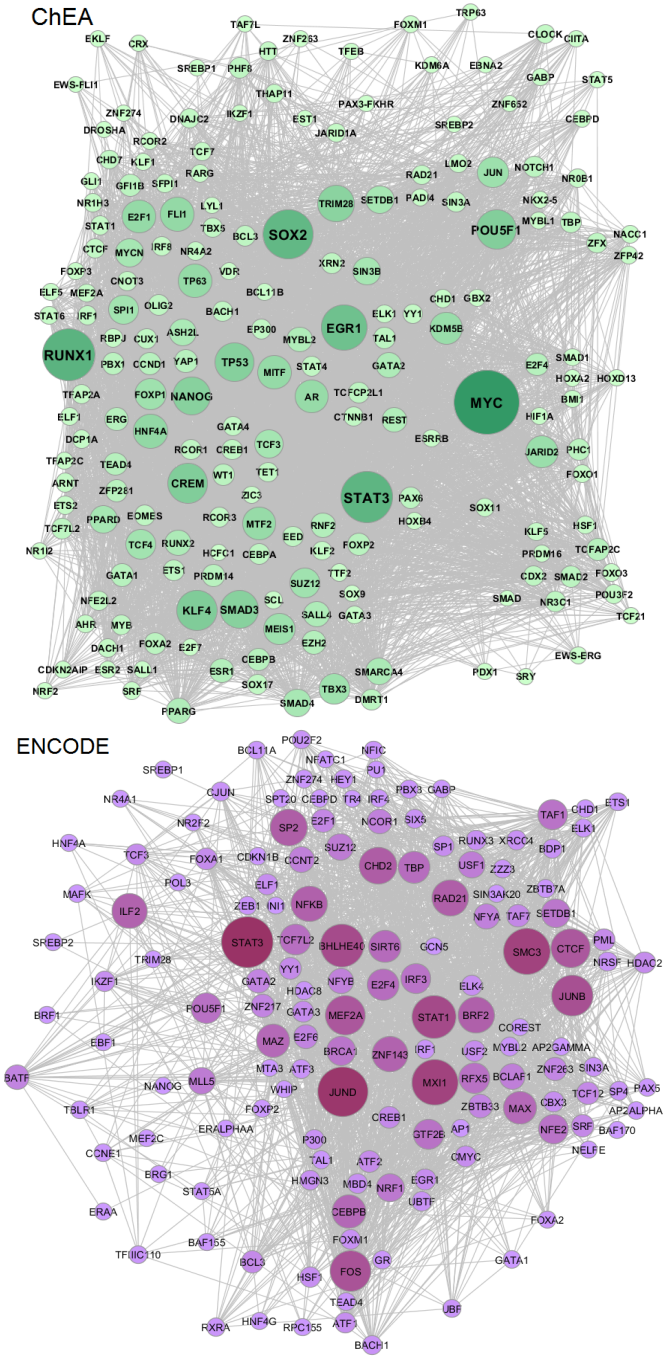


Fig. 3. Ball-and-stick representation of the transcription factor regulatory networks from ChEA (a) and ENCODE (b) based on direct regulatory interactions between transcription factors and other transcription factors. Node size and color represents connectivity degree.

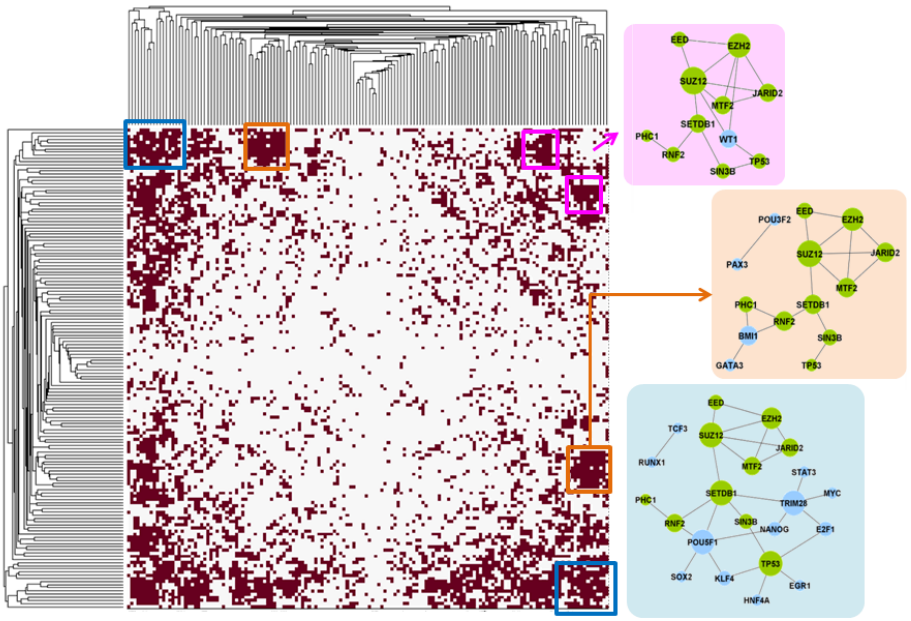


Fig. 4. Hierarchical clustering heatmap of the ChEA adjacency matrix based on direct targets. Direct protein-protein interactions connecting transcription factors that form clusters on the heatmap are highlighted for several clusters.

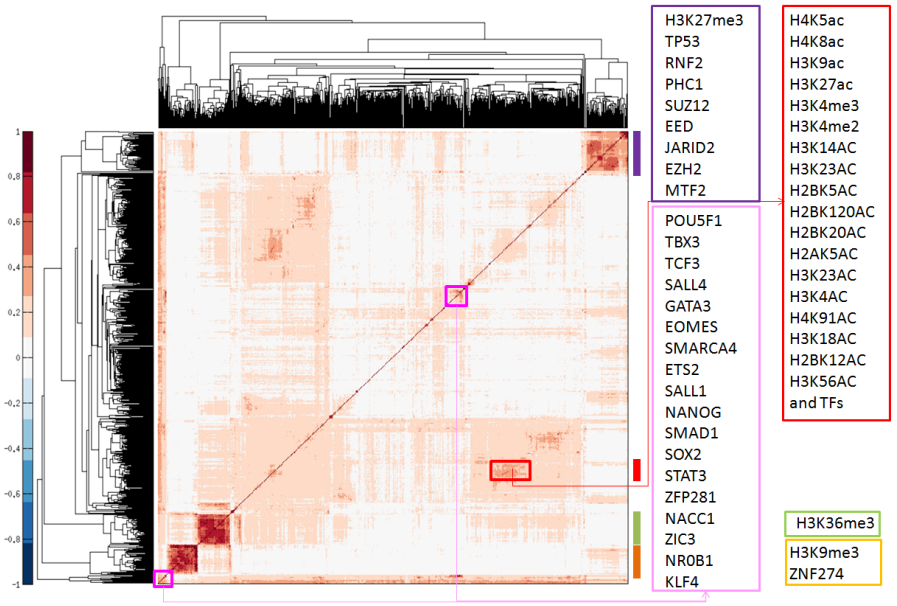


Fig. 5. Hierarchical clustering of all three gene set libraries: ChEA, ENCODE and histone modifications. Experiments were clustered based on their target-set similarity.

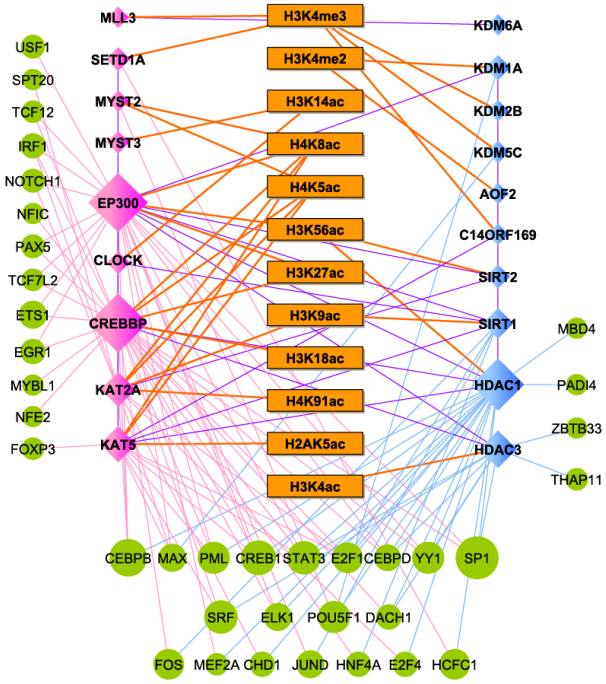


Fig. 6. Protein-protein interactions (PPI) between histone modifying enzymes and transcription factors in one of the clusters from Figure 5. Orange squares represent histone modifications, green circles are transcription factors, and pink and blue diamonds are histone modifying enzymes.

coefficients of 0.15 and 0.09 respectively. These clustering coefficients are much higher than the clustering observed for random networks of the same size. Since the density of links in these networks is very high, not much of the network structure is revealed by the ball-and-stick visualization. However, the hubs of the networks clearly stand out. These hubs are likely master regulators or well-studied factors. Interestingly BHLHE40 (also called DEC1, HLHB2, SHARP-2, STRA13, and STRA14) a less known transcription factor appears as a hub in the ENCODE network. This is because this factor is a target of many other factors as well as a regulator of many of the other profiled factors. It is possible that this transcription factor has an important global role: integrating information from many transcription factors and acting as a master regulator of many factors. This central role may have been yet mostly overlooked.

Next, we combined all three gene-set libraries into one heatmap that cluster the entries based on similarity of gene-list content (Fig. 5). Euclidean distance of the Jaccard score was used as a similarity measure and the linkage type was average. Multiple experiments forming clusters are highlighted and listed on the right. In general the histone modification experiments cluster together by modification type and the transcription factors form their own clusters. Focusing on one cluster that contains a mixture of transcription factors and histone modification experiments, we

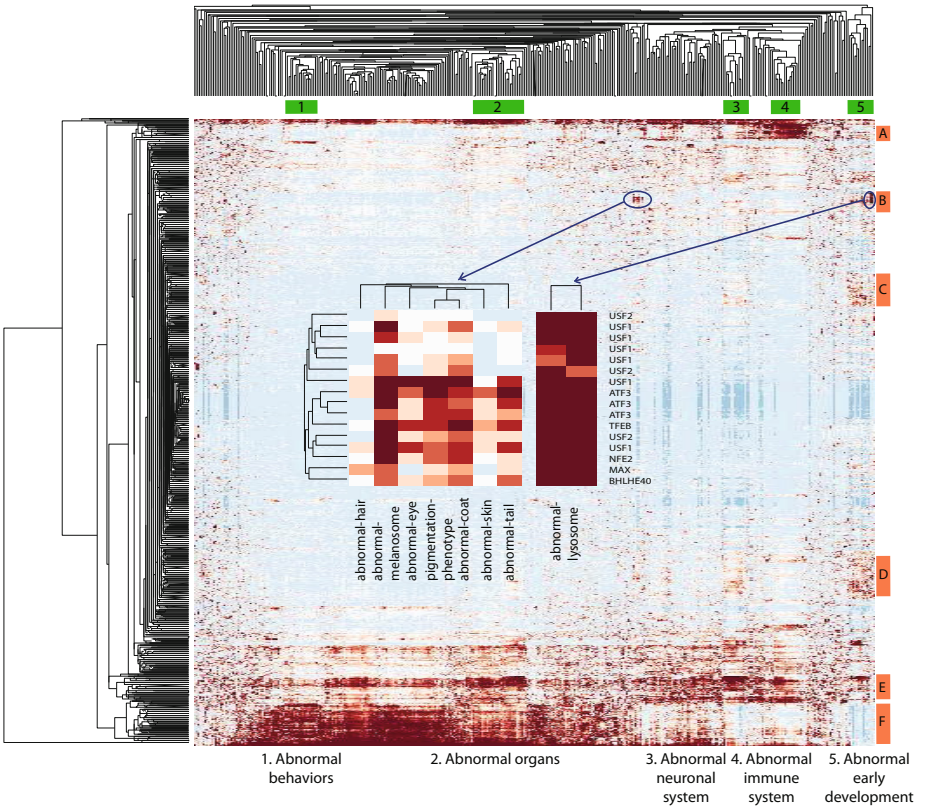


Fig. 7. Hierarchical clustering of gene sets causing abnormal phenotypes in mice and enriched transcription factors and/or histone modifications. Generalized descriptions of mouse phenotype clusters are listed at the bottom, with numbers matching to the corresponding clusters.

mapped all the known protein-protein interactions that connect the factors, as well as the protein interactions between these factors and the known enzymes that remove or add the chemical modification group on the histone from the cluster (Fig. 6). Links connect the enzymes associated with these modifications were created by a literature search and the other links represent known experimentally determined protein-protein interactions collected from available online databases [14]. We find that some factors interact more exclusively with the enzymes that remove the marks, while other factors interact more exclusively with the enzymes that add the mark. Moreover, two writers: EP300 and CREBBP and one remover HDAC1 are highly connected. This could be a literature bias since these factors are well-studied, but also can represent the type of complexes that are formed to regulate the same subset of genes. Finally, we performed enrichment analysis, using all three gene set libraries: ChEA-manual, ENCODE and histone modifications, on the entire gene set library created from the Mouse Genomics Informatics Mammalian Phenotype ontology (MGI-MP) [15].

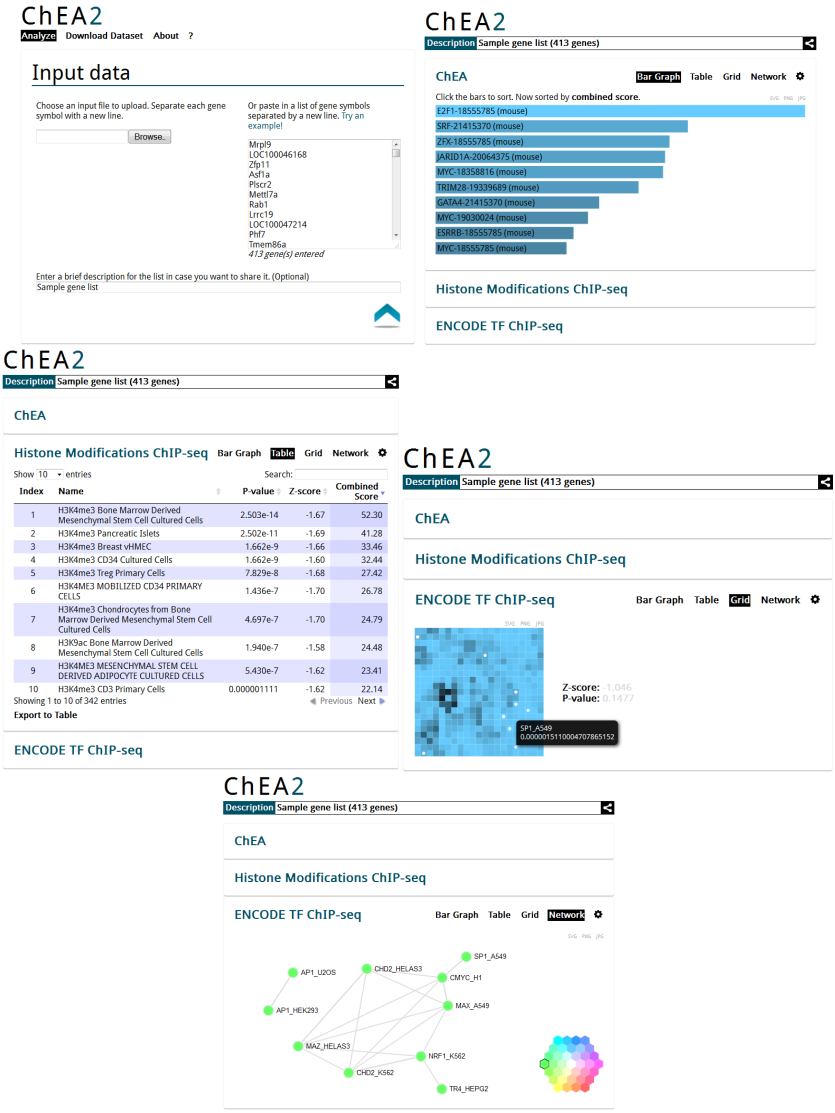


Fig. 8. Screenshots from the ChEA2 web-application. a) Users can enter lists of genes or upload a file containing lists of genes. b) Once the users submit valid gene lists, they are presented with enrichment analysis results that rank transcription factors and histone modifications as bar-graphs, tables, networks of enriched factors and canvases that highlight the enriched factors or enriched histone modification experiments.

This mapping identifies relationships between annotated single gene knockout mouse phenotypes and groups of transcriptional regulators. Genes associated with mouse phenotypes (columns) were used for enrichment analysis and the corrected p-values of each term from the three gene-set libraries: manual ChEA, ENCODE and histone

modifications (rows) were used for drawing the hierarchical clustering heatmap. Clusters of mouse phenotypes are highlighted with green labels and clusters of ChEA2 elements are labeled with orange colors (Fig. 7). For example, a group of transcription factors including GATA2, EGR1, BCL11A, FLI1, and NFKB are highly enriched for abnormal immune system as expected (Cluster 4 in Fig. 7). The mouse phenotype clusters of abnormal neuronal system (Cluster 3 in Fig. 7) and abnormal embryonic/early development (Cluster 5) share similar groups of enriched transcription factors and histone modification profiles. This may be due to the observation that neuronal specific genes are suppressed in embryonic stem cells by specific histone modifications. The map includes many potentially interesting small clusters that identify relationships between factors and phenotypes. For example, we highlight two small clusters that connect abnormal pigmentation related phenotypes and abnormal lysosomes to the transcription factors: USF1/2, ATF3, BHLHE40, MAX, NF2, and TFEB. Indeed, TFEB has a well-established role in autophagy [16] and ATF3 is known to be associated with stress responses [17]. Corre and Galibert summarized the roles of USF1/2 in various contexts including the role of these factors in pigmentation determination and stress responses [18]. While there is some related knowledge about the role of some of these factors in determining these phenotypes, our analysis suggests that these factors may work together to regulate the same sets of genes that are responsible for the induction of specific functions that lead to specific phenotypes. In addition to the preliminary meta-analysis, we performed on the three gene-set libraries ChEA2 datasets, we also developed a web-based application that provides access to the data and enables users to query their lists of differentially expressed genes (Fig. 8). The tool is implemented with HTML5 and the JavaScript library D3 and it is freely available online at http://amp.pharm.mssm.edu/ChEA2.

4 Conclusions

Here, we assembled a large compendium of ChIP-seq/chip experiments applied to mammalian cells. We began an initial meta-analysis of such complex dataset by abstracting each experiment to a gene list. Such abstraction discards much important information such as peak height and exact binding locations. Peak height and exact binding locations information are likely critical in better assessing cooperation among regulators. In addition, including such quantitative data in enrichment analyses is likely to improve the inference of regulators given sets of differentially expressed genes. Moreover, we must consider the fact that the putative target genes determined by ChIP-seq/chip contain many false positives. Integrating results from transcription factor knockdowns followed by genome-wide expression can potentially filter putative binding targets with real functional targets. Another consideration is that our network analysis is very preliminary. We did not consider measuring network properties such as connectivity distributions and identification of network motifs. Since the datasets contain a large-scale directed graphs made of transcription factors regulating other transcription factors, insights can be obtained from structural and dynamical analyses of such graphs. Regardless of these limitations, we have integrated an important large-scale dataset made from most available ChIP-X studies into a computable format and began the global analysis of this potentially useful

resource. This project paves the way for more extensive computational studies that would further unravel the transcriptional networks that are responsible for regulating the complex system of the mammalian cell. In addition, the data collection and organization, as well as the interactive light-weight web-based data visualization of ChEA2 exemplify an application from systems biology that deals with a complex interdisciplinary challenge to extract knowledge from massive datasets [19].

Acknowledgements. This work is supported in part by NIH grants R01GM098316, R01DK088541, U54HG006097-02S1, P50GM071558, and Irma T. Hirschl Career Scientist Award.

References

1. Lachmann, A., Xu, H., Krishnan, J., Berger, S.I., Mazloom, A.R., Ma'ayan, A.: ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics 26, 2438–2444 (2010)
2. Chen, L., Wu, G., Ji, H.: hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. Bioinformatics 27, 1447–1448 (2011)
3. Qin, J., Li, M.J., Wang, P., Zhang, M.Q., Wang, J.: ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. Nucleic Acids Research 39, W430–W436 (2011)
4. Lepoivre, C., Bergon, A., Lopez, F., Perumal, N., Nguyen, C., Imbert, J., Puthier, D.: TranscriptomeBrowser 3.0: introducing a new compendium of molecular interactions and a new visualization tool for the study of gene regulatory networks. BMC Bioinformatics 13, 19 (2012)
5. Qin, B., Zhou, M., Ge, Y., Taing, L., Liu, T., Wang, Q., Wang, S., Chen, J., Shen, L., Duan, X.: CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. Bioinformatics 28, 1411–1412 (2012)
6. Sun, H., Qin, B., Liu, T., Wang, Q., Liu, J., Wang, J., Lin, X., Yang, Y., Taing, L., Rao, P.K., et al.: CistromeFinder for ChIP-seq and DNase-seq data reuse. Bioinformatics 29, 1352–1354 (2013)
7. Bovolenta, L., Acencio, M., Lemke, N.: HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. BMC Genomics 13, 405 (2012)
8. Pepke, S., Wold, B., Mortazavi, A.: Computation for ChIP-seq and RNA-seq studies. Nat. Meth. 6, S22–S32 (2009)
9. Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., Peng, W.: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25, 1952–1958 (2009)
10. The ENCODE Consortium Project, An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012)
11. Chen, E.Y., Tan, C., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., Clark, N.R., Ma'ayan, A.: Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128 (2013)
12. Tan, C., Chen, E.Y., Dannenfelser, R., Clark, N.R., Ma'ayan, A.: Network2Canvas: Network Visualization on a Canvas with Enrichment Analysis. Bioinformatics 29(15), 1872–1878 (2013)

13. Clark, N., Dannenfelser, R., Tan, C., Komosinski, M., Ma'ayan, A.: Sets2Networks: network inference from repeated observations of sets. BMC Systems Biology 6, 89 (2012)
14. Berger, S., Posner, J., Ma'ayan, A.: Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. BMC Bioinformatics 8, 372 (2007)
15. Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E.: The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res. 40(1), D881–D886 (2012)
16. Decressac, M., Mattsson, B., Weikop, P., Lundblad, M., Jakobsson, J., Björklund, A.: TFEB-mediated autophagy rescues midbrain dopamine neurons from α-synuclein toxicity. Proc. Natl. Acad. Sci. U S A 110, E1817–E1826 (2013)
17. Hai, T., Wolfgang, C.D., Marsee, D.K., Allen, A.E., Sivaprasad, U.: ATF3 and stress responses. Gene Expr. 7(4-6), 321–335 (1999)
18. Corre, S., Galibert, M.: Upstream stimulating factors: highly versatile stress-responsive transcription factors. Pigment Cell Res. 18(5), 337–348 (2005)
19. Holzinger, A.: On Knowledge Discovery and interactive intelligent visualization of biomedical data - Challenges in Human–Computer Interaction & Biomedical Informatics. In: Helfert, M., Francalanci, C., Filipe, J. (eds.) Proceedings of the International Conference on Data Technologies and Application, Rome DATA 2012, Setubal (PT), pp. 3–16. SciTec Press (2012)

On the Prediction of Clusters for Adverse Reactions and Allergies on Antibiotics for Children to Improve Biomedical Decision Making

Pinar Yildirim[1], Ljiljana Majnarić[2], Ozgur Ilyas Ekmekci[1], and Andreas Holzinger[3]

[1] Department of Computer Engineering, Faculty of Engineering & Architecture, Okan University, Istanbul, Turkey
{pinar.yildirim,oekmekci}@okan.edu.tr
[2] School of Medicine, University J.J. Strossmayer Osijek, 31 000, Osijek, Croatia
ljiljana.majnaric@hit.t-com.hr
[3] Institute for Medical Informatics, Statistics & Documentation Medical University Graz, A-8036 Graz, Austria
andreas.holzinger@medunigraz.at

Abstract. In this paper, we report on a study to discover hidden patterns in survey results on adverse reactions and allergy (ARA) on antibiotics for children. Antibiotics are the most commonly prescribed drugs in children and most likely to be associated with adverse reactions. Record on adverse reactions and allergy from antibiotics considerably affect the prescription choices. We consider this a biomedical decision problem and explore hidden knowledge in survey results on data extracted from the health records of children, from the Health Center of Osijek, Eastern Croatia. We apply the K-means algorithm to the data in order to generate clusters and evaluate the results. As a result, some antibiotics form their own clusters. Consequently, medical professionals can investigate these clusters, thus gaining useful knowledge and insight into this data for their clinical studies.

Keywords: Adverse reactions and allergy (ARA), knowledge discovery, biomedical data mining, k-means algorithm.

1 Introduction

Antibiotics are the drugs most commonly prescribed to children and are most likely to be associated with allergic and adverse reactions [1],[2],[3],[4]. A reaction to a drug is considered an allergic reaction if it involves an immunologic reaction to a drug. It may occur in the form of immediate or non-immediate (delayed) hypersensitivity reactions. Immediate reactions are usually mediated with IgE antibodies (often elevated in persons with inherited susceptibility to allergic diseases, called atopy), whereas non-immediate reactions can be mediated with several other immune mechanisms [5]. The clinical manifestations of antibiotic allergy include skin reactions (varying from local and mild general to severe general reactions),

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 431–445, 2013.

organ-specific reactions (most commonly occuring in the form of blood dyscrasias, hepatitis and interstitial nephritis) and systemic reactions (usually corresponding with anaphylaxis) [5]. Many reactions to drugs mimic symptoms and signs of the allergic reactions, although being caused with non-immunologic mechanisms. In many cases, also, pathologic mechanisms remain completely unclear. This is the reason why these reactions are often considered together and commonly named adverse reactions and allergy (ARA) [6]. This term is especially appropriate for use in primary health care setting, where patients who had experienced ARA on antibiotics have rarely been referred to testing. Moreover, diagnostic tests are limited and are only standardized for penicillin allergy [6].

Antibiotic classes with higher historical use have been shown to have higher allergy prevalence [7]. Published papers on frequency, risk factors and preventability of this medical problem in the general population, and especially in children, are scarce. Available data implicate female sex, frequent use, older age, insufficient prescribing strategy and monitoring of prescribed medications, as the primary factors accounting for higher prevalence of ARA on antibiotics among adults. Similar data for children are completely absent [8].

The aim of this study is to explore hidden knowledge in the survey data extracted from health records on adverse reactions and allergy on antibiotics in children in the town of Osijek, Eastern Croatia. We plan to obtain some serious and useful information in electronic health records that are not easily recognized by researchers, clinicians and pharmaceutical companies.

2 Related Work

There have been several studies performed for knowledge discovery on drug adverse events associations. Kadoyama et al. searched the FDA's AERS (Adverse Event Reporting System) and carried out a study to confirm whether the database could suggest the hypersensitivity reactions caused by anticancer agents, paclitaxel, docetaxel, procarbazine, asparaginase, teniposide and etoposide. They used some data mining algorithms, such as proportional reporting ratio (PRR), the reporting odds ratio(ROR) and the empirical Bayes geometric mean (EBGM) to identify drug-associated adverse events and consequently, they detected some associations [9].

Tsymbal et al. investigated antibiotics resistance data and proposed a new ensemble machine learning technique, where a set of models are built over different time periods and the best model is selected. They analyzed the data collected from N.N. Burdenko Institute of Neurosurgery in Russia and the dataset consisted of some features such as: patient and hospitalization related information, pathogen and pathogen groups and antibiotics and antibiotic groups. Their experiments with the data show that dynamic integration of classifiers built over small time intervals can be more effective than the best single learning algorithm applied in combination with feature selection, which gives the best known accuracy for the considered problem domain [10].

Lamma et al. described the application of data mining techniques in order to automatically discover association rules from microbiological data and obtain alarm rules for data validation. Their dataset consists of information about the patient such as sex, age, hospital unit, the kind of material (specimen) to be analyzed (e.g., blood, urine, saliva, pus, etc.), bacterium and its antibiogram. They applied the Apriori algorithm to the dataset and developed some interesting rules [11].

3 Methods

The Study Population and Data Sources

The study was done on the population of 1491 children (769 children of the school age, 7-18 years old, the rest of the preschool age), all patients in the same Health Center in the town of Osijek, Eastern Croatia, cared for by a family physician and a primary pediatrician teams.

Data were extracted from the health records of these children. Knowledge of risk factors for ARA on antibiotics in children are scarce. In making a choice for data collection, a co-author physician used personal knowledge on factors influencing the immunologic reactions together with information from the studies on risk factors for allergic diseases in children [12],[13],[14],[15],[16],[17],[18],[19]. Data extraction, from the patients health records, was guided by a multi-item chart, in an advance prepared by this co-author. In addition, parents of children recorded on ARA on antibiotics were interviewed by telephone, on a family history of ARA on antibiotics and other allergic and chronic diseases, in which pathogenesis, in a great part, immunologic mechanisms are involved. Data were summarized.

Registered information on ARA on antibiotics was found in health records of 46 children, out of a total of 1491 children screened, implicating the overall prevalence of ARA on antibiotics of 3,15%. However, higher prevalence was found in children of the school age (4,9%), then in those of the preschool age (1,1%), data probably reflecting the cummulative incidence rates with age. When the incidence data were however estimated, it has been shown that ARA on antibiotics, in our study population, can be expected to occur predominantely in preschool age (33/46 cases, 71,1%).

Of registered ARA events, almost all were mild-moderate skin reactions. Only one case was in need for hospitalization (a18-year-old girl, treated with the combination of amoxicillin and clavulonic acid). All data, including descriptions of ARA events (upon which classification of severity reaction was made) and diagnoses of diseases, were based on the native physicians` records.

Clustering Analysis by k-means Algorithm

Cluster analysis is one of the main data analysis method in data mining research.The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another ~~a~~nd are dissimilar to the objects in other clusters. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing and biomedical research.

In this study, we use k-means algorithm to survey results on adverse reactions and allergy (ARA) on antibiotics in children. The k-means algorithm is a type of partitioning algorithm and is simple and effective. The k-means algorithm is widely used due to easy implementaion and fast execution. The algorithm takes the input parameter, *k*, and partitions a set of *n* objects into k clusters so that the resulting intra-cluster similarity is high but inter-cluster similarity is low. Cluster similarity is measured with regard to the mean value of the objects in a cluster, which can be viewed as the cluster's center of gravity. The pseudo code of a k-means algorithm is as follows:

1. arbitrarily choose *k* objects as the initial cluster centers
2. repeat
3. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster
4. update the cluster means, i.e., calculate the mean value of the objects for each cluster
5. until no change.

The distance from the mean values of the objects in the clusters are calculated by using distance measures. The Euclidean distance is one of the common distance measures and is defined as the square root of the squared discrepancies between two entities summed over all variables (i.e., features) measured. For any two entities A and B and k=2 features, say X1 and X2, d_{ab} is the length of the hypotenuse of a right triangle. The square of the distance between the points representing A and B is obtained as follows:

$$d_{ab} = [\ (X_{a1}-X_{b1})^2 + (X_{a2} - X_{b2})^2\ ]^{1/2}\ [20].$$

4 Experimental Results

We selected samples from the survey results and created a dataset. Table 1 lists the antibiotics used in the dataset. The dataset consists of 26 attributes and 42 instances (Table 2 and Table 3). The k-means algorithm was used to explore some hidden clusters in the dataset. WEKA 3.6.8 software was used. WEKA is a collection of machine learning algorithms for data mining tasks and is an open source software. The software contains tools for data pre-processing, classification, regression, clustering, association rules and visualization [21], [22].

K-means algorithm requires the number of clusters (*k*) in the data to be pre-specified. Finding the appropriate number of clusters for a given dataset is generally a trial and error process made more difficult by the subjective nature of deciding what 'correct' clustering. The performance of a clustering algorithm may be affected by the chosen value of *k*. Reported studies on k-means clustering and its applications usually do not contain any explanation or justification for selecting particular values for *k* [23].

Table 1. Type of antibiotics used in survey

Type of antibiotics	
Short name used in the dataset	**Full name and information**
ampicilin	ampicilin
cef& pen	cefalosporins & penicillin
pen&klav	penicillin & amoxicillin+clavulanic acid
klav	amoxicillin+clavulanic acid - a broad-spectrum
azitrom	azithromycin - a macrolide group
cef	cefalosporins - a broad-spectrum
fenoksi	fenoksimetil penicillin - per os penicillin, a narrow-spectrum
cefuroks	cefuroxime - the second generation of cefalosporins - a broad-spectrum
pen	penicillin
sulfa	sulfamethoxazole
eritrom	erythromycin - a macrolide antibiotic of an older generation

The k-means algorithm implementation in many data-mining or data analysis software packages requires the number of clusters to be specified by the user. To find a satisfactory clustering result, usually, a number of iterations are needed where the user executes the algorithm with different values of k [23]. In order to evaluate the performance of simple k-means algorithm in our study, two test modes were used, training set and percentage split (holdout method).The training set refers to a widely used experimental testing procedure where the database is randomly divided into k disjoint blocks of objects, then the data mining algorithm is trained using k-1 blocks and the remaining block is used to test the performance of the algorithm, this process is repeated k times. At the end, the recorded measures are averaged. It is common to choose k=10 or any other size depending mainly on the size of the original dataset.

In percentage split (holdout method), the database is randomly split into two disjoint datasets. The first set, which the data mining system tries to extract knowledge from called training set. The extracted knowledge may be tested against the second set which is called test set, it is common to randomly split a dataset under the mining task into 2 parts and has 66% of the objects of the original database as a training set and the rest of objects as a test set. Once the tests were carried out using our dataset, results were collected and an overall comparison was conducted [24].

Table 2. The attributes used in the dataset (1-17)

No	Attribute	Description	Type
1	Age	The patient`s age	Numeric
2	Age of ARA	Age when the allergic/adverse reaction on antibiotics occurred	Numeric
3	Type of antibiotic	Generic name of the antibiotic by which the allergic reaction was provoked	Nominal
4	Severity reaction	The clinically graded allergic/adverse reaction	Ordinal
5	Age of the 1st antibiotic use (y)	Age when the first antibiotic was used	Numeric
6	Other allergic disease (skin)	Does a child have some other allergic disease? (manifestation on the skin)	Nominal (Yes, No)
7	Other allergic disease (rhinitis)	Does a child have some other allergic disease? (in the form of allergic rhinitis)	Nominal (Yes, No)
8	Other allergic disease (bronchitis)	Does a child have some other allergic disease? (in the form of obstructive bronchitis)	Nominal (Yes, No)
9	Other allergic disease (asthma)	Does a child have some other allergic disease? (in the form of asthma)	Nominal (Yes, No)
10	Blood test on allergy - IgE	Have the antibodies of the IgE type (which usually raise in allergic diseases) been measured?	Nominal (Positive, Negative)
11	Perinatal disorders	Disorders occurring during delivery and the first hours after the birth	Nominal (Yes, No)
12	The child birth order	Born as the first, or the second, etc., child in order	Ordinal
13	Severe respiratory disease	A respiratory disease which is severe enough to be a life frightening (e.g. laryngitis, pneumonia)	Nominal (Yes, No)
14	Age of severe respiratory disease	Age when some type of severe respiratory disease occurred	Numeric
15	Otits media	Otits media	Nominal (Yes, No)
16	Age of otitis media	Age when otitis media occurred	Numeric
17	Other infections	Had there been some other infection before the allergic/adverse reaction on antibiotics occured?	Nominal (Yes, No)

Table 3. The attributes used in the dataset (18-26)

No	Attribute	Description	Type
18	Other infections (the number of episodes)	How many episodes of infections had there been before the allergic/adverse reaction on antibiotics occurred?	Nominal
19	Varicella	Did the varicella infection occur?	Nominal (Yes, No)
20	Age of varicella	Age when varicella infection occurred	Numeric
21	Hospitalization <2y of age	Hospitalization in the very early childhood	Nominal (Yes, No)
22	Number of infections per year	An average number of infections per year in a particular child, independently on when the allergic/adverse reaction on antibiotics occurred	Numeric
23	Antibiotic exposure before ARA	How many times antibiotics had been prescribed before the allergic/ adverse reaction on antibiotics occurred?	Ordinal
24	Family history on ARA	Family history on allergic/adverse reactions on antibiotics	Nominal(Positive, Negative)
25	Allergic diseases in family	Have there been other allergic diseases in family members?	Nominal(Yes, No)
26	Chr diseases diseases in family	Have there been other chronic diseases in family members?	Nominal (Yes, No)

Table 4. Evaluation of cluster analysis with training set mode

K value	Number of iterations	Within cluster sum of squared errors	Runtime(Seconds)
2	3	459.114	0.01
3	4	444.553	0.01
4	3	430.279	0.01
5	5	415.160	0.01

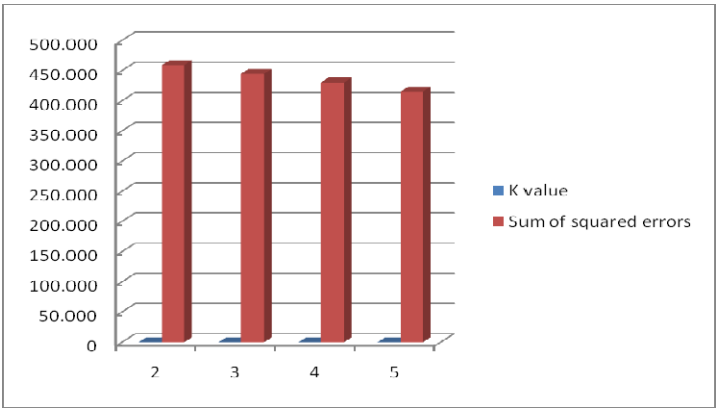


Fig. 1. Sum of squared errors and k values with training set mode

Table 5. Evaluation of cluster analysis with percentage split test set mode

K value	Number of iterations	Within cluster sum of squared errors	Runtime(Seconds)
2	3	459.114(66%)	0
	3	293.226(34%)	
3	4	444.553(66%)	0.01
	5	279.846(34%)	
4	3	430.279(66%)	0.01
	5	264.258(34%)	
5	5	415.160(66%)	0.01
	2	248.785(34%)	

We also tried different number of clusters (2<=*k*<=5) for each test mode and we observed the results of number of iterations, sum of squared errors and runtime. Sum of squared error (SSE) is an evaluation measure that determines how closely related are objects in a cluster. The formula of SSE is as follows:

$$\text{SSE}=\sum_{i}\sum_{x\in Ci} dist^2(x, m_i)$$

C_i is the *j*th cluster, m_i = centroid of cluster C_i (the mean vector of all the data points in C_j), and $dist(\mathbf{x}, \mathbf{m}_j)$ is the distance between data point **x** and centroid $\mathbf{m}_j$ [25].

The results after analysis are described in Table 4 and 5. We compared the results of the number of clusters obtained by simple k-means algorithm and we found that greater number of clusters produced smaller sum of squared errors. For example, when k value is 2 which is default in Weka, sum of squared error is 459.114, on the other hand, when k value increased to 4, new value of sum of squared error is 430.279(Fig 1). Table 6 shows clusters with training set mode and with k=4.

Table 6. Clusters obtained by k-means algorithm with training set mode and k=4

Attribute	Cluster1	Cluster2	Cluster3	Cluster4
Age	14.5	14.0	16.0	16.0
Age of ARA	5.0	6.0	0.4	<1
Type of antibiotic	cef&pen	pen	fenoksi	ampicilin
Severity reaction	skin	skin	skin	skin
Age of the 1st antibiotic use (y)	5.0	<1	<1	<1
Other allergic disease (skin)	yes	no	no	yes
Other allergic disease (rhinitis)	no	no	no	no
Other allergic disease (bronchitis)	no	no	no	yes
Other allergic disease (asthma)	no	no	no	no
Blood test on allergy - IgE	positive	positive	positive	positive
Perinatal disorders	yes	no	no	yes
Birth order	1	1.379	1.2828	1.2685
Severe respiratory disease	yes	no	no	Yes
Age of severe respiratoy disease	1.5	6.0	6.0	6.0
Otitis media	yes	yes	no	Yes
Age of otitis media	9.0	<1	<1	<1
Other infections	yes	yes	yes	Yes
Other infections (the number of episodes)	1X	2X	2X	2X
Varicella	yes	no	yes	no
Age of varicella	7.5	3.0	3.0	3.0
Hospitalization <2y of age	no	no	no	no
Number of infections per year	3-4X	2-3X	2-3X	3X
Antibiotic exposure before ARA	2X	1X	1X	1X
Family history on ARA	negative	negative	negative	positive
Allergic diseases in family	no	no	no	no
Chr diseases diseases in family	no	no	no	no

The results of the k-means algorithm revealed some patterns in the survey data and four clusters were generated (Table 6). According to the results, some types of antibiotics form their own clusters such as cef&pen, pen, fenoksi and ampicilin.

Medical researchers and clinicians can consider and explore these patterns to create some medical ideas.

5 Discussion

This is a collaborative study of an interdisciplinary team, composed of informaticians and a physician (a GP). The role of a physician was in forming a research question and data collection and in providing comments on health-related issues.

An overall frequency of ARA on antibiotics of 3,15% was observed. Rarely available data for the paediatric population indicates the overall incidence of 9,35% in hospitalized children and 1,46% in outpatients [26]. Many factors can affect the variation in the frequency of this disorder, including the children age (as shown in our paper), the natural distribution of risk factors in the population, the types of antibiotics prescribed, the custom of ARA recording and physicians` education on both, symptoms and mechanisms of ARA and antibiotic prescription [27].

When all four clusters in parallel were put into consideration, some general rules, in regard to ARA in children, coud be observed. As the first, there were two time peaks of the ARA occurrence: in the year of birth and in the late pre-school age (around 5-6 y). In older children with ARA, the causing antibiotics were classified as with higher historical use (penicillin).

Some common characteristics of children with ARA might include: 1) predisposition to allergic disorders (positive IgE blood test), 2) however, not manifested with allergic respiratory diseases (hay fever and asthma). This connection should be taken into account even if it is known that allergic diseases show the time-dependent occurrence during the childhood (the so-called “allergic march”, manifested as a progression of atopic diseases from eczema to asthma), for reason that the current age of cases with ARA corresponds with adolescence (14-16 year). These results seem contrary to what is known from the early studies, that atopic subjects do not show higher incidence of penicillin allergy, in comparison to the general population [6]. It cannot, in fact, be known, from our results, whether atopy in children can also increase their predisposition for ARA on drugs (especially on antibiotics other than penicillin), or whether, on the contrary, early antibiotic exposure increases the risk for atopic diseases, as postulated traditionally [13]. Or, these results may be only due to the confounding effects, consequently to the predominant use of β-lactam antibiotics in children. Namely, undesirable reactions on these antibiotics are known as being predominantly caused by allergic mechanisms, usually mediated with IgE antibodies [6]. Nevertheless, these results can direct future prevention strategies, mainly by means of preserved prescriptions of antibiotics in children with increased IgE antibodies.

Other constant and common features of children with ARA include: 3) at least one episode of infections (other than respiratory infections, also including otitis media) experienced before the time of ARA occurrence, as well as an early antibiotic use (in the first year of life). These results might be reflective of the immune system disturbation, in the early childhood, which can increase the chance for both, ARA on antibiotics and infections. Also, there are information that some infections can serve

as a promoting factor, by ensuring conditions for the immune reaction on a drug to start, which otherwise could not be the case [5]. In addition to these explanations, the second result might also implicate the increased risk for ARA to occur, through the negative effect of an early antibiotic use on the commensal intestinal flora and the subsequent impairments of the immune system development [17], [18].

Some additional factors, found to commonly occur in children with ARA, include: 4) frequent infections (defined as two or more times per year), reflecting poor hygiene, or the immune system dysfunction, and 5) low antibiotic pre-exposure counts (1-2 times), indicating sensitizing reaction as the possible mechanism of ARA. In accordance to the latter, it is commonly known that patients usually develop allergic reactions when reexposed to an antibiotic [6].

When clusters 3 and 4, representing an early onset of ARA (during the first year of life), were compared to each other, somewhat different patterns were obtained, probably indicating different mechanisms between ARA on ampicilin (a broad-spectrum antibiotic) and fenoksimetilpenicilin (a narrow-spectrum penicillin for an oral use). Otherwise, these antibiotics share the common structure, that of the β-lactam antibiotic group, also sharing some common features [5].

In regard to ampicilin, other allergic diseases, including skin eczema and obstructive bronchitis (both disorders occurring early along the course of the "allergic march"), may contribute to the onset of ARA. These results are likely to support the hypothesis, already presented above, about the common pathogenetic background of both, atopic diseases and ARA on antibiotics, in children. As an alternative explanation of this connection, evidence has been provided by many clinical studies, although not consistently, that antibiotic exposure in early infancy is likely to increase the risk for childhood atopy [13], [17]. This inconsistency in knowledge gained on this issue, might be the consequence of the different behavior of otherwise similar substances, such as in our study the case with fenoksimetil penicillin (cluster 3) and ampicillin (cluster 4). The unfavourable drug reaction, in ampicilin risk group (cluster 4), according to our results, might also be supported with the existence of perinatal disorders, implicating immunodeficiency and obstacles in the postnatal immune system development. In numerous studies, conducted to-date, an attention has not been payed to the importance of these very early developmental disturbances. Furthermore, our results also indicate that the occurrence of otitis media in early life (in some reports considered as the complication of influenza virus infection and, as such, the manifestation of the immune system dysfunction) can also be considered as a contributing factor for the early onset of ARA on ampicillin (cluster 4). This risk group, in contrast to the comparative one, for the time of onset (cluster 3), was also prone to the development of severe respiratory disease, although with the onset later in life (at six age), further indicating immunodeficient disorders. When positive family history on ARA is added to this risk group (cluster 4), this all together indicates that a set of inherited and acquired immune system disorders can be important for the occurrence of ARA on this broad-spectrum antibiotic.

Some elements of this pattern, associated with ARA on ampicillin (cluster 4), can be recognized as a part of the cluster describing ARA on cefalosporins (cef&pen, cluster 1), another broad-spectrum group of antibiotics. These elements, overlapping

between the two clusters, include perinatal disorders and severe respiratory disease, although here, the severe respiratory disease preceded (and probably contributed to) the onset of ARA (cluster 1). The combined cef&pen ARA event probably means allergic cross-reaction that may occur between penicillin and cephalosporins of the older generation [6].

Also, it is interesting to observe that two very similar antibiotics, from the common penicillin groups (clusters 2 and 3), have gained much of the similarity in their risk factors patterns.

These results, indicating multiple factors clustered within distinct patterns, each of them specifically associated with a particular risk group (or an antibiotic), are similar to the results of the studies on the association of an early antibiotics use and the occurrence of allergic diseases later in the childhood. According to these studies, a complex cause/outcome model should be formed, in order to make conclusions on this issue, and it is not possible to achieve by analyzing only one, or even a few risk factors [13], [14], [18].

All these factors, extracted from the health records and selected within four clusters, reflect patients` (children`s) clinical and pathophysiological features. We can speculate that the reason why ARA on some other antibiotics, also listed above, have not been presented with a cluster, might be the need for different clinical parameters selection, those ones not recorded in the health records. Alternatively, some other factors could be responsible for ARA, such as, for example, differences in pharmacodynamic mechanisms of drug action. In contribution to this latter explanation, very low ARA rates for macrolide antibiotics have been reported [5].

Results of this study have confirmed some relatively known facts about ARA in children, including the influence of early life infections and antibiotic prescriptions, as well as the predomination of allergic mechanisms underlying ARA, mostly mediated with IgE antibodies. The nature of the association between atopy and ARA in children, also important for understanding childhood allergic diseases, remain to be elucidated in the future. In fact, our results indicate that this association might be important only for early ARA onset (in the first year of life) and for a particular antibiotic used. The main contribution of this paper is in the results clearly showing for the first time that only a cluster of factors can explain ARA, specifically for a particular children group, or an antibiotic.

Results of this study can further be utilized for planning future research on this issue. They can also be useful when preparing recommendations for antibiotics prescription and to guide the standardized health data record. Merely an increase in awareness of physicians on risk factors for ARA in children can be sufficient to change their attitudes towards antibiotics prescription. Computer-based tools would be helpful in many aspects when managing these issues, especially by means of the possibility for systematic data recording and data modeling, suitable for the purpose of prediction and risk factors identification. Also important would be the drug allergy alert and prescription support systems, as well as programs for education promotion [28], [29].

We analyze health records created in a health center in East Croatia to explore new knowledge for adverse reactions and allergy (ARA) on antibiotics in children.

The broad application of business enterprise hospital information systems amasses large amounts of medical documents, which must be reviewed, observed, and analyzed by human experts. There is need for techniques which enable the quality-based discovery, the extraction, the integration and the use of hidden knowledge in those documents [30]. Human-Computer Interaction and Knowledge Discovery along with Biomedical Informatics are of increasing importance to effectively gain knowledge, to make sense out of the big data. In the future, we can combine these fields to support the expert end users in learning to interactively analyze information properties thus enabling them to visualize the adverse reactions and allergy (ARA) on antibiotics data[31].

6 Conclusion

Biomedical research aims to explore new and meaningful knowledge to provide better healthcare [32]. Adverse reactions and allergy (ARA) from antibiotics in children is an important research issue for the medical domain. In this study, we focused on knowledge discovery for this problem and perform a study based on data mining to predict clusters in the survey data extracted from health records of children in Eastern Croatia.

We used computational techniques and then applied k-means algorithm to the dataset to generate some clusters which have similar features. Our results highlight that some type of antibiotics form different clusters. Medical researchers and pharmaceutical companies can utilize and interpret our results. Despite that our study has some limitations, for example we have small dataset consisting of 42 instances, we hope that we can extend the dataset and apply data mining algorithms on it in the future.

In conclusion, we believe that our study can be good example on data mining for adverse reactions and allergy (ARA) from antibiotics in children.

Acknowledgements. We thank the CD-ARES reviewers for their thorough review and helpful comments to further improve our paper.

References

[1] Langley, J., Halperin, M., Allergy, S.: to antibiotics in children: perception versus reality. Pediatric Infectious Disease Notes 13(3), 160–163 (2002)

[2] Kramer, M.S., Hutchinson, T.A., Flegel, K.M., Naimark, L., Contardi, R., Leduc, D.G.: Adverse drug reactions in general pediatric outpatients. J. Pediatr. 106, 305–310 (1985)

[3] Menniti-Ippolito, G., Raschetti, R., Da Cas, R., Giaquinto, C., Cantarutti, L.: Active monitoring of adverse drug reactions in children. Italian Paediatric Pharmacosurveillance Multicenter Group. Lancet 355, 1613–1614 (2000)

[4] Cirko-Begovic, A., Vrhovac, B., Bakran, I.: Intensive monitoring of adverse drug reactions in infants and preschool children. Eur. J. Clin. Pharmacol. 36, 63–65 (1989)

[5] Thong, B., Update, Y.-H.: on the management of antibiotic allergy. Allergy Asthma Immunol. Res. 2(2), 77–86 (2010)

[6] Robinson, J., Hameed, L., Carr, T., Practical, S.: aspects of choosing an antibiotic for patients with a reported allergy to an antibiotic. Clin. Infect. Dis. 35, 26–31 (2002)
[7] Macy, E.T., Poon, K.Y.: Self-reported Antibiotic Allergy Incidence and Prevalence: Age and Sex Effects. The American Journal of Medicine 122(8), 778 (2009)
[8] Krizmanic, V., Majnaric, L.: Adverse reactions and allergy on antibiotics in children. In: SouthCHI 2013, Sloenia (2013)
[9] Kadoyama, K., Kuwahara, A., Yamamori, M., Brown, J.B., Sakaeda, T., Okuno, Y.: Hypersensitivity Reactions to Anticancer Agents:Data Mining of the Public Version of the FDA Adverse Event Reporting System, AERS. Journal of Experimental & Clinical Cancer Research 30(93), 1–6 (2011)
[10] Tsymbal, A., Pechenizkiy, M., Cunningham, P., Puuronen, S.: Dynamic Integration of Classifiers for tracking concept drift in antibiotic resistance data. Information Fusion 9, 56–68 (2008)
[11] Lamma, E., Manservigi, M., Mello, P., Nanetti, R.F., Storari, S.: The automatic discovery of alarm rules for the validation of microbiological data. In: IDAMAP 2001, London, UK (2001)
[12] Harris, J.M., Mills, P., White, C., Moffat, S., Taylor, A.J.N., Cullinan, P.: Recorded infections and antibiotics in early life: associations with allergy in UK children and their parents. Thorax 62, 631–637 (2007)
[13] Johnson, C.C.J., Ownby, D.R., Alford, S.H., Havstad, S.L., Williams, K., Zoratti, E.M., Peterson, E.L., Joseph, C.L.M.: Antibiotic exposure in early infancy and risk for childhood atopy. J. Allergy Clin. Immunol. 115, 1218–1224 (2005)
[14] Halken, S.: Prevention of allergic disease in childhood: clinical and epidemiological aspects of primary and secondary allergy prevention. Pediatr. Allergy Immunol. 15(suppl. 16), 9–32 (2004)
[15] Headley, J., Northstone, K.: Medication administered to children from o to 7.5 years in the Avon longitudinal study of parents and children (ALSPAC). Eur. J. Clin. Pharmacol. 63(2), 189–195 (2007)
[16] Hawkins, N., Golding, J.: A survey of the administration of drugs to young infants. Br. J. Clin. Pharmacol. 40(1), 79–82 (1995)
[17] Bremner, S.A., Carey, I.M., DeWilde, S., Richards, N., Maier, W.C., Hilton, S.R., Strachan, D.P., Cook, D.G.: Early-life exposure to antibacterials and the subsequent development of hayfever in childhood in the UK: case-control studies using the general practice research database and the doctors‘ independent network. Clin. Exp. Allerg. 33, 1518–1525 (2003)
[18] Thomas, M., Custovic, A., Woodcock, A., Morris, J., Simpson, A., Murray, C.S.: Atopic wheezing and early life antibiotic exposure a nested case-control study. Pediatr. Allergy Immunol. 17, 184–188 (2006)
[19] Kozyrskyj, A.L., Ernst, P., Becker, A.B.: Increased risk of childhood asthma from antibiotic use in early life. Chest 131, 1753–1759 (2007)
[20] Han, J., Micheline, K.: Data mining: concepts and techniques. Morgan Kaufmann, San Francisco (2001)
[21] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
[22] WEKA: Weka 3: Data Mining Software in Java, `http://www.cs.waikato.ac.nz/ml/weka` (last access: April 18, 2013)
[23] Pham, D.T., Dimov, S.S., Nguyen, C.D.: Selection of k in k-mens clustering. Journal of Mechanical Engineering Science 219, Part C (2005)

[24] Tiwari, M., Bhai Jha, M.: Enhancing the performance of data mining algorithm in letter image recognition data. International Journal of Computer Applications in Engineering Sciences 223(4946), 217–220 (2012)
[25] Tan, P., Steinbach, M., Kumar, V.: Introduction to data mining. Lecture Notes (2004)
[26] Gomes, E., Fonseca, R., Drug, J.: allergy claims in children: from self-reporting to confirmed diagnosis. Clin. Experiment Allergy 38, 191–198 (2007)
[27] Impicciatore, P., Choonara, I., Clarkson, A., Provasi, D., Pandolfini, C., Bonati, M.: Incidence of adverse drug reaction in paediatric in/out-patients: a systemic review and meta-analysis of prospective studies. Br. J. Clin. Pharmacol. 52, 77–83 (2001)
[28] Madle, G., Kostkova, P., Weinberg, J.: Bugs and drugs on the Web: changes in knowledge of users of a web-based education resource on antibiotic prescribing. Journal of Antimicrobial Chemotherapy 63(1), 221–223 (2009)
[29] Farrell, et al.: Computer games to teach hygiene: an evaluation of the e-Bug junior game. Journal of Antimicrobial Chemotherapy 66(suppl. 5), v39–v44 (2011)
[30] Holzinger, A., Yildirim, P., Geier, M., Simonic, K.-M.: Quality-based knowledge discovery from medical text on the Web Example of computational methods in Web intelligence. In: Pasi, G., Bordogna, G., Jain, L.C. (eds.) Qual. Issues in the Management of Web Information. ISRL, vol. 50, pp. 145–158. Springer, Heidelberg (2013)
[31] Holzinger, A.: On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human–Computer Interaction & Biomedical Informatics. In: Conference on e-Business and Telecommunications (ICETE 2012), Rome, Italy, IS9–IS20 (2012)
[32] Holzinger, A., Simonic, K.M., Yildirim, P.: Disease-disease relationships for rheumatic diseases: Web-based biomedical textmining and knowledge discovery to assist medical decision making. In: 36th International Conference on Computer Software and Applications COMPSAC, pp. 573–580. IEEE, Izmir (2012)

A Study on the Influence of Semantics on the Analysis of Micro-blog Tags in the Medical Domain

Carlos Vicient and Antonio Moreno

Department of Computer Science and Mathematics, Universitat Rovira i Virgili,
Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA) research group,
Av. Països Catalans, 26. 43007, Tarragona, Catalonia, Spain
{carlos.vicient,antonio.moreno}@urv.cat

Abstract. One current research topic in Knowledge Discovery is the analysis of the information provided by users in Web 2.0 social applications. In particular, some authors have devoted their attention to the analysis of micro-blogging messages in platforms like Twitter. A common shortcoming of most of the works in this field is their focus on a purely syntactical analysis. It can be argued that a proper semantic treatment of social tags should lead to more structured, meaningful and useful results that a mere syntactic-based approach. This work reports the analysis of a case study on medical tweets, in which the results of a semantic clustering process over a set of hashtags is shown to provide much better results than a clustering based on their syntactic co-occurrence.

Keywords: Semantic similarity, tags, co-occurrence, clustering, micro blogging.

1 Introduction

In the last years the World Wide Web has moved from a collection of static pages that were created by experts and read by users to a very dynamic environment in which users are not only consuming information but also providing it (the so-called Social Web or Web 2.0)[1]. Many researchers in Artificial Intelligence are already working in the next step (the Social Semantic Web or Web 3.0 [2]), in which intelligent autonomous agents should be able to understand the semantics behind all this user-generated data in order to solve automatically complex processes like open information extraction, information retrieval, question answering, automated reasoning, etc. [3–5].

In the current Social Web users generate and distribute many different kinds of data in a wide variety of formats, from pure text to pictures, audios, videos, weblogs, etc. [6]. In order to facilitate the access to these data, many Web 2.0 applications allow users to attach them some kind of keywords, usually called tags, which provide information on the main topics related to a certain item. In some limited cases these tags may belong to well-defined domain taxonomies or ontologies, providing an accurate indexing of the tagged elements. However, in most cases uses may freely add any textual keywords, giving rise to unstructured folksonomies. An important

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 446–459, 2013.

research problem is the design and development of tools that allow users to visualise, explore and interact with these enormous unstructured repositories of user-tagged data (HCI) and to discover and extract meaningful knowledge from them (KDD)[7]. Current tools usually rely on information retrieval techniques to find relevant pieces of information about a certain topic on such a huge volume of generated data and also on clustering methods to sort and classify the relevant documents according to their similarity. In this regard, in the last years some researchers have proposed ontology-based text clustering [8] or tag similarity [9, 10] methods. This classification process allows a posterior filtering process that keeps only those sets of documents that are more related to the user's interests, which may then finally be shown to the user applying advanced HCI visualization techniques. However, despite the great number of works in the field, it can be argued that there is still a lack of semantically-based classification and visualization mechanisms on Social Web data.

Some researchers are focusing on the analysis of the information provided by users in social networks. Micro-blogging services are one of the Web 2.0 kinds of applications that are attracting more attention. The most successful one is probably Twitter, that has around 200 million regular users that generate 500 million tweets per day [11]. A tweet is a free text with a maximum length of 140 characters. Users may tag a tweet using the so-called hashtags, which are strings of characters that begin with the "#" symbol.

As will be commented in the next section, many researchers have focused on the analysis of hashtags in order to perform complex knowledge-based tasks on a corpus of tweets [12, 13]. However, most of the scientific contributions have focused almost exclusively on a purely syntactic analysis of the content of tweets, including their hashtags. For instance, some works have proposed to discover the main topics in a corpus of tweets by clustering them taking into account the syntactic co-occurrence of hashtags. In this paper we want to argue that a semantic analysis of hashtags, based on the use of well-known ontology-based semantic similarity measures, should lead to more meaningful and relevant results that a merely syntactic analysis. We have tested this research hypothesis by clustering a set of medical tweets using both syntactic co-occurrence and semantic similarity measures, obtaining a more appropriate classification of the tweets in the second case, as we expected.

The rest of the paper is structured as follows. The next section provides a brief background on related work on the analysis of tweets, showing the current predominance of syntactic methods. Section 3 describes a new mechanism that permits to make a semantic analysis of the hashtags contained in a set of tweets, by associating them to the concepts of an ontology and then employing ontology-based similarity measures to compute the relatedness between hashtags. In section 4 the case of study using a set of medical tweets is presented, explaining the analysed data set and the syntactic and semantic clustering techniques applied on them. The following section presents and comments the results of the study, and the last section closes the paper with some final comments and the presentation of future lines of work.

2 Related Work

There have been many works in which sets of tweets are analysed for different purposes (e.g. visualization of clusters of similar tweets [14, 15], recommendation of hashtags [16], sentiment analysis [17, 18], detection of events or common topics [19, 20], clustering of tweets [21, 22], etc.). However, in most cases it can be observed that the treatment of the tweet components (mainly the words contained in the body of the tweet and its hashtags) is purely syntactic. Kywe et al [16] aim to recommend hashtags that could be associated to a particular tweet, by considering those hashtags employed by similar users (those that use the same hashtags) or used in similar tweets (those that contain the same words). Doan et al. [19] analyse sets of tweets in order to track the evolution of influenza, and they propose to filter those tweets that contain certain pre-defined words as hashtags. Russell et al. [20] analyse sets of tweets related to the Energy domain, and they define a notion of "semantic similarity" between tweets based on the co-occurrence of their terms. Bhulai et al. [14] also consider the co-occurrence of words within tweets to make clusters of tweets for visualization purposes. Teufl and Kraxberger [23] represent a tweet with a set of terms (nouns, adjectives, verbs and hashtags) and use the co-occurrence between them to define a weighted graph that can be analysed to obtain the "semantic pattern" associated to each tweet. Mathiesen et al. [24] build graphs of terms, where the edges are weighted by their co-occurrence, and study the communities present in these graphs. Veltri [21] also uses the co-occurrence between words to classify tweets related to the Nanotechnology domain. The lack of a semantic treatment of the content of the tweets is the main shortcoming of all these approaches.

Some authors have focused on the analysis of the co-occurrences between the hashtags that appear on a set of tweets, but they also seem to take a purely syntactic point of view. Wang et al. [17] define a sentiment analysis method based on different mechanisms of sentiment propagation in graphs that basically take into account the co-occurrence of hashtags. Ozdikis et al. [22] cluster hashtags by considering the co-occurrence between the hashtags and the words that appear in the tweets. Pöschko [15] also considers the co-occurrence between hashtags to group together related tweets for visualization purposes. Again, all these approaches consider only the co-occurrence of the strings that define the tags, without trying to understand their meaning to make a more thorough and semantically coherent treatment.

The main aim of this paper is to show, via an illustrative case of study in the medical domain, that clustering hashtags using a proper semantic similarity measure (supported by a domain ontology) provide a much more coherent result than the use of a purely syntactic co-occurrence metric.

3 Semantic Tag Similarity

As mentioned before, one of the main characteristics of social tagging is that users can freely annotate contents without any restriction in their choice of tags. Those tags usually lack any form of explicit organization and normalization, giving rise to unstructured folksonomies. This fact produces, as a consequence, several problems when using tags in retrieval tasks and in their classification. One of them is that

different tags might have been used for the same concept (synonymy), which makes it difficult to find all items relevant for a certain concept. Another one is that the same tag can have different meanings in different contexts (polysemy). Just to name another one, a purely syntactic analysis of tags is unable to show whether two tags have some kind of relatedness (e.g. "Religious Building" is more general than "Cathedral", and "Cathedral" is more related to "Abbey" than to "Restaurant"). To overcome these problems, it is necessary to have methods that can find tags that are conceptually related. In other words, a semantic tag similarity measure is needed in order to compare two tags at the conceptual level (i.e. a metric to measure how similar is the meaning of two lexically different tags).

Another characteristic of this kind of social tagging is the fact that tags may be acronyms, named entities, a composition of different terms or even invented words. This diversity presents a challenge in the discovery of the conceptual meaning of the tag. The linkage between the term and its meaning (a concept in a background knowledge structure, for instance a domain ontology) is what we call semantic annotation, and solving this issue is required to be able to define semantic similarity measures for tags. In this work WordNet [25], a semantic lexicon for the English language that models and semantically interlinks more than 100,000 concepts referred to by means of English textual labels, is used as the reference background knowledge structure into which tags are mapped. WordNet is organised in synsets, which are useful to solve problems like synonymy. Moreover, it also introduces hierarchies of concepts which can be used to identify related concepts.

The rest of this section aims to propose a methodology to link tags with its meaning (with WordNet elements) and to calculate the semantic similarity between them. Notice that tags are not necessarily contained in WordNet. So, the basic idea is to find the mapping between tags and WordNet concepts and, after that, to apply well-known semantic measures to compare two different tags associated with WordNet concepts. Thus, the presented methodology is divided in two main parts: the semantic annotation and the calculation of the semantic similarity.

3.1 Semantic Annotation

The aim of this first stage is to analyse a tag and associate it, if possible, with a WordNet concept. In case the tag is composed of multiple words, the system generalises it by dropping sequentially the leftmost terms until a matching is found (e.g., if the tag is "Gothic Cathedral" and this expression is not found in WordNet, the system will look for "Cathedral").

Algorithm 1 shows the procedure applied to calculate the similarity between two tags. The input parameters are the two tags to be compared and the semantic measure to be used to compare two WordNet elements. The function *getCandidates*, explained later, returns a list of possible concept candidates to be matched with the tag within WordNet. This list only contains the tag itself if it matches directly with a WordNet concept. However, due to the nature of the hashtags commonly used in Twitter, there may be many cases in which hashtags are not found directly in WordNet and a more complex analysis is needed to find out the WordNet concept that should be associated to the hashtag (in fact, in some cases it may turn out to be impossible to find an

appropriate matching between a hashtag and any concept). After that, if both candidate lists are not empty, the semantic similarity between the tags is computed by the *calculateSimilarity* function, which is also described below.

```
Sem_Sim(tag1, tag2, measure){
  candTag1 := getCandidates(tag1)
  candTag2 := getCandidates(tag2)
  if (candTag1 || candTag2) == null
    return 0.0
  return calculateSimilarity(candTag1, candTag2, measure)
}
```

Algorithm 1. Calculates the semantic similarity between tags "tag1" and "tag2", considering a given "measure" of semantic similarity between WordNet concepts

In order to find possible semantic annotations for a given tag, the first step is to check whether it appears directly as a WordNet concept. If the tag is found, then the annotation is direct. However, as pointed out before, there are many different kinds of tags that do not appear in WordNet directly. To overcome this problem, the proposed methodology relies on Wikipedia as an external knowledge base. This procedure can be seen in the pseudo-code of Algorithm 2.

```
getCandidates(tag){
  wTag := getWNConcept(tag)
  candidates <- wTag
  if wTag == null
    candidates := getWikipediaCandidates(tag)
  return candidates
}
```

Algorithm 2. Given a tag, returns a lists of possible WordNet concepts with which it could be annotated

Wikipedia is a well-known free online encyclopaedia which contains more than 30 million articles that have been written collaboratively by volunteers all around the world. In all the knowledge-based tasks associated to the analysis of Natural Language it has been assumed in the last years that Wikipedia is the premier and more comprehensive repository of textual knowledge, due its enormous breadth and the quality of its collaborative-based contents [26]. Its articles are not limited to the standard <concept, definition> structure of a dictionary, but they can describe just about anything one can imagine. Thus, Wikipedia entries have a much wider scope than WordNet concepts, permitting to find for example acronyms, named entities, different lexicalizations of the same concept, etc. Moreover, Wikipedia articles are loosely classified by means of a hierarchy of Wikipedia categories. Each of them defines the essential characteristics of a certain topic, allowing readers a mechanism to browse and quickly find sets of related pages.

In the semantic annotation mechanism, when a tag has not been found in WordNet, it is looked up on Wikipedia. If there is an entry for the tag, all the associated categories are retrieved. A category is treated as a phrase and it is proposed as an annotation candidate only if its head (the main noun of the description of the category, extracted by a natural language parser) matches with a WordNet concept. This process is shown in the high level description of the function *getWikipediaCandidates* in Algorithm 3. So, eventually, all the proposed candidates are WordNet concepts. A tag will not have any associated candidate concepts only if it is not found either in WordNet or in Wikipedia, or if the categories found in Wikipedia do not match with any concept in WordNet.

```
getWikipediaCandidates (tag){
  wikiCandidates := null
  if existsWikiEntry(tag)
    auxCategories := getCategoriesFromWiki(tag)
    forall cat ∈ auxCategories
      mainNoun := getNN(cat)
      auxCat := getWNConcept(mainNoun)
      if auxCat != null
        wikiCandidates <- auxCat
  return wikiCandidates
}
```

Algorithm 3. Used to extract and process Wikipedia categories that will work as candidates to be annotated with WordNet concepts

3.2 Semantic Similarity

At this point there is a list of WordNet concepts (obtained directly or via Wikipedia categories) associated to each of the two tags to be compared. In order to establish the degree of semantic relatedness of the tags, the similarity between all the pairs of candidates (one from each tag) is calculated. This similarity is calculated with a function that computes the semantic similarity measure between WordNet concepts, which is given as parameter by the user. This function could be any of the well-known ontology-based similarity metrics described in the literature. The final similarity between the tags is the maximum of the similarities between the associated WordNet concepts. If the tags are associated to a certain domain of knowledge (for instance, the medical domain considered in the next section), it may be argued that calculating the similarities between all the pairs of candidates and taking the maximum one solves, in an indirect way, the problem of disambiguating the correct sense of the tag. The idea is that, even if the terms that we are comparing are polysemic, each of them will have a "medical" sense, and these domain-related senses will be the ones with a higher similarity. This process is shown in the pseudo-code depicted in Algorithm 4.

```
calculateSimilarity(candidates1, candidates2, measure){
  simMax := 0
  forall cat1 ∈ candidates1
    forall cat2 ∈ candidates2
      sim := measure(cat1, cat2)
      if(sim >= sim_max)
        simMax := sim
  return simMax
}
```

Algorithm 4. Maximises the similarity between all the pairs of candidates of the given lists.

4 Case of Study: Data Set of Medical Tweets

The underlying research work concerns the study of the influence of the use of ontology-based semantic similarity measures on the analysis of a corpus of tweets. In this paper we present a preliminary study, conducted over a set of medical tweets on the cancer disease. As will be argued in the next section, the results of this study confirm the intuition on the improvement of the results with respect to the usual syntactic analysis. This section describes the obtainment and treatment of the data set, whereas the next section provides a more detailed analysis of the results of the comparative study between syntactic and semantic clustering of tags.

The first step was the extraction of a dataset of medical tweets from the Symplur Website[1]. This dataset is composed of approximately 5000 tweets that are strongly related with "cancer" (according to Symplur), dated from October 31st 2012 to January 11th 2013. After a first filtering pre-process, we obtained a set of 1086 different hashtags used in this dataset. The graph on Fig. 1 depicts a distribution of the number of hashtags per tweets, where the y-axis represents the total number of tweets that contain exactly the number of hashtags represented in the x-axis. Notice that there are 2530 tweets tagged only with one hashtag and just one tweet containing 12 distinct hashtags.

The first step of the analysis is the semantic annotation, in which hashtags are linked to WordNet concepts. 409 hashtags (37.6 %) are found directly in WordNet. When Wikipedia categories are used to support the annotation process, as described in the previous section, the system is able to annotate 815 hashtags (75.0%). Therefore, this indirect annotation route permits to double the number of annotated hashtags. However, it has to be noted that 25% of the hashtags cannot be annotated even with the help of Wikipedia. An area of future work is to make a more detailed analysis of the hashtags so that this percentage of unannotated tags can be lowered; moreover, another field of future work is the thorough analysis of the quality of the annotation, both in its direct and indirect routes.

[1] www.symplur.com. Last access: May 24th, 2013.

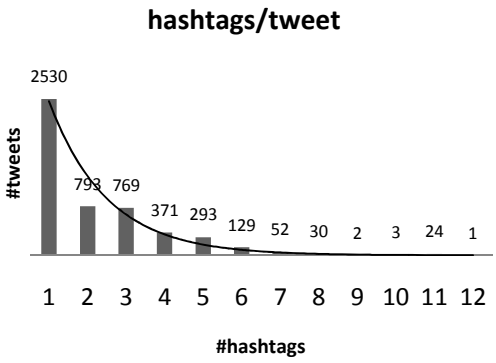


Fig. 1. Distribution of hashtags per tweet

A common way of analysing a set of tweets is to take their hashtags and partition them in a set of non-overlapping classes, so that all the hashtags in the same class (cluster) are supposed to be heavily related, whereas those belonging to different classes should be associated to different topics [15, 17, 22]. In this study we have compared two different ways of clustering the set of hashtags of the tweet corpus:

- The first clustering process is purely syntactic. The measure of similarity between two hashtags used by the clustering algorithm is based on the number of tweets in which both hashtags co-occur. The more frequently two hashtags appear, the more related they are supposed to be.
- The second clustering process is grounded on the use of a domain ontology that permits to measure the actual semantic similarity between two hashtags.

The pseudo-code of the algorithm used in this study is shown in Algorithm 5. The two input parameters of the algorithm are k (the number of classes to be generated in the clustering process) and the main topic of the extracted tweets (which in this study is "Cancer"). First, we obtain the set of tweets related to the input topic. After that, a filtering method is applied. In this step duplicates (e.g. re-tweets of the same tweet by different users) are removed, and word-breaking techniques are used in order to split hashtags composed by more than one word. The next step, described in the following subsections, is the construction of a similarity matrix between hashtags, which is used in the final step by a clustering algorithm to obtain the set of clusters of hashtags.

```
Hashtag_Clustering (k, topic){
  TW ← getTweets(topic)
  TW ← filterTweets(TW)
  Mtw ← generateSimilarityMatrix(TW)
  Ctw ← hierarchicalClustering(k, Mtw)
}
```

Algorithm 5. Algorithm used to perform the clustering of hashtags

4.1 Syntactic Clustering Based on Hashtag Co-occurrences

In order to generate the symmetric similarity matrix used by the clustering process two steps are necessary. First, a hashtag co-occurrence matrix (as shown in Eq. 1) is constructed, where n represents the total number of hashtags, c_{ij} is the number of co-occurrences between hashtag i_{th} and hashtag j_{th} within the dataset, and c_{ii} contains the times that the hashtag i_{th} appears in the whole dataset. Notice that c_{ij} has the same value that c_{ji} , fact that implies that the order in which the hashtags appear in the set of tweets is not relevant.

$$\mathrm{C}_n = \begin{bmatrix} c_{11} & c_{12} & c_{1n} \\ \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & c_{nn} \end{bmatrix} \mid c_{ij} \in \mathbb{N} \tag{1}$$

Afterwards, the matrix has to be normalised so that all its values are between 0 and 1, since the clustering method used in this study needs either a similarity or a dissimilarity matrix as input data. To do so, each element of the matrix is normalised ($\forall i \in [1,n] \forall j \in [1,n]\, c_{ij} = NormalizedM(i,j)$) where $NormalizedM(i,j)$ is the function shown in Eq. 2. This function assigns the value 1 to each element of the diagonal of the matrix (a hashtag is equal to itself). The co-occurrence between two different hashtags is normalised by dividing it by the minimum number of their individual appearances. The rationale of this approach is that two hashtags will be considered highly similar if, in most of the tweets in which one of them appears, the other hashtag also appears. For instance, if hashtag A appears in 100 tweets, hashtag B appears in 500 tweets, and they appear together in 80 tweets, their similarity will be very high (80/100=0.8), since B appears in 80% of the tweets containing A. Other more restrictive ways of normalising the matrix values, considering for instance the maximum or the average of the individual appearances, could have also been considered.

$$\mathrm{NormalisedM(i,j)} = \begin{cases} 1, & \mathrm{i = j} \\ \frac{\mathrm{c_{ij}}}{\mathrm{MIN(c_{ii}, c_{jj})}}, & \text{otherwise} \end{cases} \tag{2}$$

In Fig. 1 it was shown that more than half of the analysed tweets only contain 1 hashtag, and only around 18% of them have 4 or more hashtags. Therefore, the number of co-occurrences between hashtags is relatively small; moreover, the pairs of co-occurrent hashtags are likely to appear in different tweets. Therefore, there are many cells in the co-occurrence matrix with a 0 value (near to 86%), for all those pairs of hashtags that do not appear together in any of the input tweets (i.e. only about 14% of pairs of hashtags co-occur). With this co-occurrence based similarity measure, these pairs of hashtags are considered to be totally dissimilar.

4.2 Clustering Based on Semantic Similarities

In this section the aim is the same than in the previous section, the construction of a normalised similarity matrix between hashtags. However, in this case we want to define an ontology-based semantic similarity matrix. Let S_n (Eq. 3) be the mentioned matrix, where n represents the total number of hashtags and s_{ij} is the semantic

similarity between hashtag i_{th} and hashtag j_{th} , calculated with the SEM_{sim} function (Algorithm 1), so that $\forall i \in [1,n] \forall j \in [1,n]\ c_{ij} = SEM_{sim}(i,j,measure)$.

$$\mathrm{S}_n = \begin{bmatrix} s_{11} & s_{12} & s_{1n} \\ \cdots & \cdots & \cdots \\ s_{n1} & s_{n2} & s_{nn} \end{bmatrix} \mid s_{ij} \in [0,1] \quad (3)$$

In the experiment reported in this section Wu and Palmer's similarity measure [27] has been used to estimate the semantic alikeness between words by mapping them to WordNet concepts and computing the number of semantic links separating them. This measure gives a value between 0 and 1, so it does not require any further normalisation. As a result, terms are classified according to their semantic similarity.

It is worth noting that the S_n matrix does not have as many null values as the normalised C_n matrix. The reason is that C_n relies on direct co-occurrences among hashtags, and most of them do not tend to co-occur at all, whereas S_n is constructed with the results of the SEM_{sim} function, which is applied to WordNet elements. Thus, two very different hashtags will probably have a very low semantic similarity, but not a null one.

5 Analysis of the Results

The final part of the study consists in clustering the set of 815 annotated hashtags in two different ways, considering the C_n and S_n normalised similarity matrixes, and comparing the results. The intuition is that the semantic clustering should be more meaningful and better structured than the one based on syntactic co-occurrence. In this case, the parameter k has been fixed to 99 clusters (to obtain clusters with an average low number of elements, around 8, which can be easily managed and analysed) and the resultant clusters have been compared as follows:

1. *Centroids* are key components in many data analysis algorithms such as clustering. They basically represent a central value that minimises the distance to all the objects in the cluster. The centroid of each cluster has been calculated with the ontology-based semantically-grounded methodology presented in Martínez et al [28]. In this approach, first a set of centroid candidates is constructed, taking into account all the concepts associated to the input cluster of hashtags according to the background knowledge (i.e. WordNet). In a second step, the final centroid of each cluster is calculated as the concept candidate that minimises the average semantic distance to all concepts in the set.
2. In order to determine the internal *compactness* of each cluster (how similar they elements are) we have computed in this study the average distance to the centroid of all the elements of the cluster which are linked to WordNet concepts. Notice that, in order to calculate this distance, any well-known ontology-based semantic similarity measure like path length, Wu-Palmer, etc. [27, 29] could have been used, since all these linked elements are in the semantic level defined by the WordNet taxonomy. Budanitsky et al [30] reported a list of those measures and evaluated their performance. In this study, as mentioned before, we have used the Wu-Palmer semantic similarity measure [27]. This measure is a path length-based measure that has the advantage of being independent of corpus statistics, uninfluenced by sparse data and easy to implement.

3. The average distance of all the elements of each cluster with respect to its centroid is calculated for both approaches (co-occurrence and semantic) and the results are compared. The lower the distance, the better the cluster (i.e. the inter-homogeneity of the cluster is stronger for lower distances).

Table 1. Average distances of clusters

		% cluster elements found in WordNet							
		10%		25%		33,3%		50%	
		Cn	Sn	Cn	Sn	Cn	Sn	Cn	Sn
match > 1	%	7,1%	**52,5%**	7,1%	**51,5%**	7,1%	**49,5%**	7,1%	**39,4%**
	avg	0,25	**0,20**	0,25	**0,20**	0,25	**0,19**	0,25	**0,18**
match > 2	%	3,0%	**35,4%**	3,0%	**35,4%**	3,0%	**33,3%**	3,0%	**27,3%**
	avg	0,31	**0,21**	0,31	**0,21**	0,31	**0,21**	0,31	**0,21**
match > 3	%	2,0%	**26,3%**	2,0%	**26,3%**	2,0%	**25,3%**	2,0%	**23,2%**
	avg	0,37	**0,23**	0,37	**0,23**	0,37	**0,23**	0,37	**0,23**

Table 1 shows the results after applying the hierarchical clustering "hclust"[2] algorithm provided by the R library separately on the C_n and S_n matrixes. Columns represent the percentage of elements of the clusters that have a direct match with WordNet concepts (e.g. the first column indicates the percentage of clusters for which at least 10% of their elements have been identified as WordNet concepts). Rows correspond to the minimum number of elements of the cluster which have been directly linked to WordNet concepts (for instance, the last row refers to those clusters that have more than 3 elements which have been identified as WordNet concepts). The combination of a row and a column restricts the possible size of a cluster. If there are *N* elements in a cluster which have been found in WordNet, and the minimum required percentage of cluster elements in WordNet is *P* (given as a number between 0 and 100), the number of elements of the cluster must be between *N* and 100**N*/*P*.

If the table results are analysed, it can be observed that in all cases the percentage of clusters that meet the requisites specified above is much bigger for the semantic analysis and, moreover, their average distances are always smaller, which means that the clusters are more compact (i.e. the inter-group homogeneity is stronger). On the other hand, the fact that the percentage of clusters obtained from C_n drops quickly to very low levels (2-3%) when the minimum number of WordNet matchings is increased is due to the fact that the majority of the clusters obtained from this syntactic co-occurrence matrix are very small (from 1 to 3 elements). On the contrary, the size of the clusters obtained from the semantic similarity matrix S_n is much more homogeneous, and 23-26% of the clusters still have more than 3 elements identified as WordNet concepts. Thus, the clusters obtained in this case are much more semantically meaningful and useful than those obtained in the first case, in which there is basically a very big cluster accompanied by a large number of very small and irrelevant clusters. This fact is also shown in Fig**.** 2 which depicts the distribution of elements per cluster in a logarithmic scale.

[2] `http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html` Last access: May 24th, 2013.

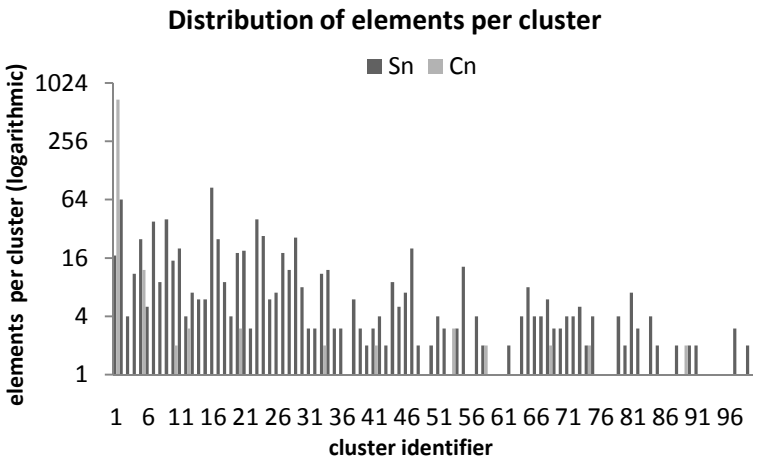


Fig. 2. Distribution of the elements per cluster in a logarithmic scale

6 Conclusions and Future Work

This study has analysed two different ways of clustering the set of hashtags that appear in a given corpus of medical tweets. A classification is based on the syntactic co-occurrence of the hashtags, whereas the other one focuses on the semantic similarity between the WordNet concepts associated to the hashtags (either directly or with the support of Wikipedia categories). The results of the case study reported in this paper seem to support the initial intuition that the common syntactic-based analysis of social tags should be replaced by more complex semantically-based treatments that provide a better structure to the knowledge extracted from Web 2.0 social applications.

There are several lines of work that can be pursued in the future. It is possible to think of more complex ways in which hashtags (especially those composed of multiple words) may be linked to WordNet concepts. The content of a tweet could be used to disambiguate the sense of the hashtag that it contains, if the hashtag corresponds to a WordNet synset with several senses (now this disambiguation is made in a more implicit way, with the pairwise comparison of all the candidate concepts associated to the two hashtags to be compared). Concerning the similarity matrix based on co-occurrences, it could be possible to normalise it using other functions rather than the minimum (recall Eq. 2). It would also be interesting to explore similarity measures based on second-order co-occurrences (two hashtags A and B could not appear together very often in a set of tweets, but they could separately co-occur quite often with another hashtag C).

Acknowledgments. This work was partially supported by the Universitat Rovira i Virgili (pre-doctoral grant of C. Vicient, 2010BRDI-06-06) and the Spanish Government through the project DAMASK-Data Mining Algorithms with Semantic Knowledge (TIN2009-11005).

References

1. O'Reilly, T.: What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software (2005)
2. Berners-Lee, T., Hendler, J.: The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 284, 34–43 (2001)
3. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Commun. ACM. 51, 68–74 (2008)
4. Fensel, D., Bussler, C., Ding, Y., Kartseva, V., Klein, M., Korotkiy, M., Omelayenko, B., Siebes, R.: Semantic web application areas. In: Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems, NLDB (2002)
5. Brill, E.: Processing natural language without natural language processing. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 360–369. Springer, Heidelberg (2003)
6. Holzinger, A., Kickmeier-Rust, M.D., Ebner, M.: Interactive technology for enhancing distributed learning: a study on weblogs. In: Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology, pp. 309–312. British Computer Society, Swinton (2009)
7. Holzinger, A.: On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human-Computer Interaction & Biomedical Informatics. In: Helfert, M., Francalanci, C., Filipe, J. (eds.) DATA. SciTePress (2012)
8. Hotho, A., Staab, S., Stumme, G.: Wordnet improves Text Document Clustering. In: Proc. of the SIGIR 2003 Semantic Web Workshop, pp. 541–544 (2003)
9. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems. In: Baumeister, J., Atzmüller, M. (eds.) LWA. Department of Computer Science, pp. 18–26. University of Würzburg, Germany (2008)
10. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
11. Hold, R.: Twitter in numbers. The Telegraph (2013), `http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html`
12. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. Presented at the (2011).
13. Weng, J., Lim, E.-P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 261–270. ACM, New York (2010)
14. Bhulai, S., Kampstra, P., Kooiman, L., Koole, G., Deurloo, M., Kok, B.: Trend visualization in Twitter: what's hot and what's not? In: Data Analytics 2012, The First International Conference on Data Analytics, pp. 43–48. IARIA, Barcelona (2012)
15. Pöschko, J.: Exploring Twitter Hashtags. The Computing Research Repository, CoRR (2011)
16. Kywe, S.M., Hoang, T.-A., Lim, E.-P., Zhu, F.: On recommending hashtags in twitter networks. In: Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., Guéret, C. (eds.) SocInfo 2012. LNCS, vol. 7710, pp. 337–350. Springer, Heidelberg (2012)

17. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: CIKM 2011, pp. 1031–1040 (2011)
18. Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., Holzinger, A.: Opinion Mining on the Web 2.0 – Characteristics of User Generated Content and Their Impacts. In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013. LNCS, vol. 7947, pp. 35–46. Springer, Heidelberg (2013)
19. Doan, S., Ohno-Machado, L., Collier, N.: Enhancing Twitter Data Analysis with Simple Semantic Filtering: Example in Tracking Influenza-Like Illnesses. In: Healthcare Informatics, Imaging and Systems Biology (HISB), pp. 62–71. IEEE Computer Society (2012)
20. Russell, M.G., Flora, J., Strohmaier, M., Poschko, J., Rubens, N.: Semantic Analysis of Energy-Related Conversations in Social Media: A Twitter Case Study. In: International Conference of Persuasive Technology (Persuasive 2011), Columbus, OH, USA (2011)
21. Veltri, G.A.: Microblogging and nanotweets: Nanotechnology on Twitter. Public Understanding of Science (2012)
22. Özdikiş, Ö., Şenkul, P., Oguztüzün, H.: Semantic expansion of hashtags for enhanced event detection in Twitter. In: The First International Workshop on Online Social Systems, WOSS (2012)
23. Teufl, P., Kraxberger, S.: Extracting semantic knowledge from twitter. In: Tambouris, E., Macintosh, A., de Bruijn, H. (eds.) ePart 2011. LNCS, vol. 6847, pp. 48–59. Springer, Heidelberg (2011)
24. Mathiesen, J., Yde, P., Jensen, M.H.: Modular networks of word correlations on Twitter. Sci. Rep. 2 (2012)
25. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). MIT Press (1998)
26. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Web Semantics 6, 203–217 (2008)
27. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics, Stroudsburg (1994)
28. Martínez, S., Valls, A., Sánchez, D.: Semantically-grounded construction of centroids for datasets with textual attributes. Knowledge-Based Systems 35, 160–172 (2012)
29. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics 19, 17–30 (1989)
30. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Comput. Linguist. 32, 13–47 (2006)

Graphic-Based Concept Retrieval

Massimo Ferri

Dip. di Matematica and ARCES, Univ. di Bologna, Italy
massimo.ferri@unibo.it

Abstract. Two ways of expressing concepts in the context of image retrieval are presented. One, Keypics, is on the side of an image owner, who wants the image itself to be found on the Web; the second, Trittico, is on the side of the image searcher. Both are based on the paradigm of human intermediation for overcoming the semantic gap. Both require tools capable of qualitative analysis, and have been experimented by using persistent homology.

Keywords: concept retrieval, image retrieval, persistent homology, size functions, keypics.

1 Introduction

A problem, which is obviously central in information retrieval, is that making a machine understand the content of an image (or any other document, for that matter) is a hard task; making it understand a concept is still tougher. There is ongoing research on extracting concepts from texts, based on occurrences (Bag-Of-Words) and on various, ingenuous stucture analysis (ontologies), and on statistics on very large data sets [6,11,17]. Similar ideas have been exported and adapted to concept extraction from images (e.g. within ImageCLEF [15] and in general in the Content-Based Image Retrieval community).

The meaning of words drifts in time. The semantic content of images undergoes even more rapid and drastic changes. The link between *signifier* and *signified* depends very much on the cultural, spatio-temporal environment, on the specific tasks of the user, even on age and gender [19]. Therefore we believe that conceptual annotation will be possible — at least for some more years — only with a human intermediation bridging the "semantic gap" [18]. This is presently performed in various forms, but mostly by textual or numerical indexing [16].

We present here two research lines developed by our Vision Mathematics group at the University of Bologna in the last few years; they are not new, but we think that they might be little known by and of some interest for the Human-Computer Interaction community; both find a place in Semiotic Engineering [14,13,12]. In both, the human intervention is essential and is expressed graphically. The first research line, "Keypics", consists of an annotation by hand-drawn sketches. It was suspended because its effectiveness depends on a social factor: for it to be successful, it ought to be adopted by a very large community. The second line, "Trittico", is based on the idea that a visual concept can be "spanned" by

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 460–468, 2013.

different images, the concept itself being the *quid* in common. It was abandoned as a commercial project but is now being revived in the paradigm of relevance feedback [10].

2 Keypics

A first way of transferring (or condensing, or stressing) concepts contained in a document, in particular an image, is by indexing. However, indexing and retrieving by words suffers from several drawbacks: the language barrier, the existence of synonims and namesakes, and above all rigidity. In fact, concepts have fuzzy and moving boundaries, but words crystallize them into discrete objects among which movement occurs by leaps. On the contrary, drawings are more dynamic. The solution we propose is indexing by *Keypics* (as opposed to key*words*). In practice, we suggest that images on the Internet should be equipped with simplified sketches representing the essentials of the images themselves (see also [5,4]). The sketches should be provided by the image owner or manager. This graphical indexing might be extended to whole Web pages.

This might be performed by use of simple drawing and processing tools, or by hand. The Keypic should represent what is felt as essential by the image owner. So it could be an outline of the relevant shapes in the image, or a symbol semantically referring to its content. E.g., the picture of a symphonic orchestra (or its home page) might be indexed by a note, meaning that the picture concerns music. Several images might be associated to the same Keypic, and more than one Keypic might be associated to the same image.

Fig. 1. A clip-art image of a toucan and its Keypic

The idea of using iconic or graphical metadata is surely not new. The most common example is perhaps that of road signs; although some text often accompanies them, road signs are generally conceived as neutral with respect to

language. Their shape is not necessarily related in a semantic way to the message they carry: It is mostly conventional, although the choice of the shape may be dictated by psychological considerations. Another noticeable situation in which shapes substitute or at least accompany a textual indication is sports: as far as we know, the universally accepted signs for the different specialities, were designed for the 1964 Olympics in Tokyo for overcoming the obvious linguistic problems.

We propose that the icons for picture indexing should be simple, easy to draw, easy to process; they should either refer to the geometric aspects of the indexed pictures, or to their semantic contents, or both. They should preferably be expressed with a compact, standard code. They should be plastic, in the sense that they should not be limited to any pre-defined set. They should be, in terms of an image, as synthetic, meaningful and free as keywords are in general use. Actually, they would be superior to keywords, in that they would not suffer from the linguistic barrier, they would allow much more freedom of expression, they would be less severely affected by errors.

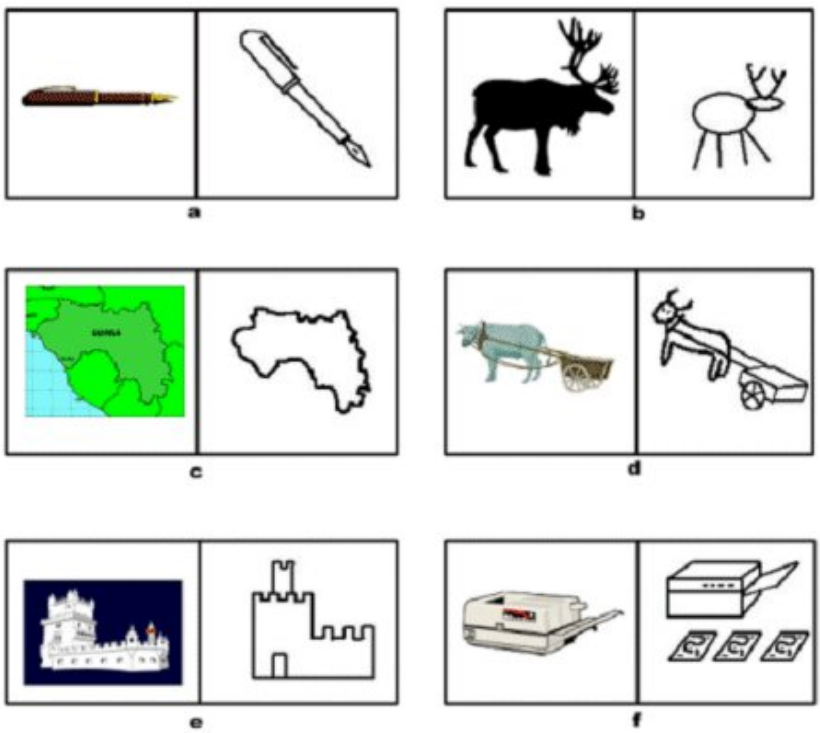


Fig. 2. Different strategies in drawing Keypics

We are aware that a likely and easy solution, which we consider deeply wrong, would be the creation of a fixed set of icons. This would imply that only a limited — even if wide — set of ideas might be conveyed. Moreover, users should depend on the choices of external authorities and maybe even on the claims of copyright owners. Updating would be necessary and frequent, with all problems related to version compatibility.

For these reasons, we stress the importance of leaving the highest freedom of expression to the image owner. This does not mean that stereotypes should be avoided; only that they should not be imposed. Actually, we believe that attractors will arise spontaneously by imitation. As naturally as new words are continually created and subjected to the natural selection of use, new Keypics would arise first in special circles, then possibly spread out to a wider community. They would be left free to appear, evolve (in a far smoother way than words) and eventually disappear.

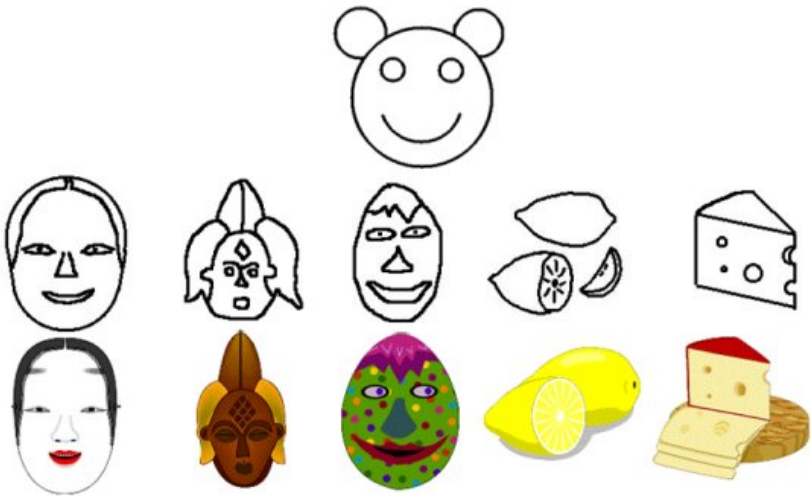


Fig. 3. A (partly) successful query

Another advantage of the plasticity we propose, lies in the rendering of morphological (and possibly semantic) nuances. As an example, the image owner who uploads a toucan image should be so provident as to detail the large beak as in Figure 1. Then, the image would be retrieved both by a user looking for birds, and (with greater priority) by one strictly interested in toucans.

How to process such drawings? The solution we adopted in an experiment is the geometrical-topological tool of Persistent Betti Numbers in degree zero (also called Size Functions) [7,8,1,3,9]. They are modular shape descriptors particularly apt to capturing qualitative aspects of images.

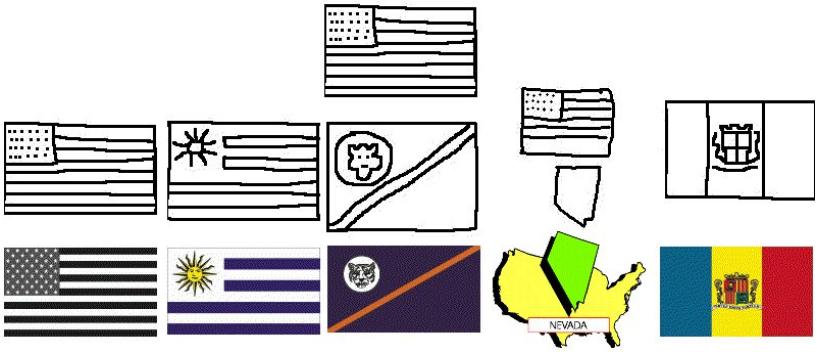


Fig. 4. An unexpected output

Seven nonprofessional draftsmen were given templates chosen within very heterogeneous pictures of a commercially available clip–art collection; the stated aim was to depict the essentials of the given template, not to reproduce it accurately. A standard drawing program was used by all of them, endowed with standard tools as free–hand, straight–line or ellipse drawers, thresholding and edge detection. A set of 494 drawings resulted of it, all of a standard size, all black on white.

We were surprised by the very heteogeneous strategies adopted. Some drew a fairly accurate imitation as in Figure 2a. Sometimes the imitation was very rough (Figure 2b); in other cases (e.g. in Figure 2c) the use of an edge detector was evident. Some draftsmen thought it necessary to stress details (Figure 2d), or to ignore them (Figure 2e), but sometimes even to add nonexisting ones (Figure 2f).

We obtained rather satisfactory results in term of precision-recall curves in a set of retrieval experiments, but what we would like to stress here is that the semantic gap was actually filled effectively by the Keypic producers (see e.g. Figure 3). Sometimes this happened in unexpected ways. It was, e.g., the case of a query with the USA flag, where the map of Nevada popped up, because the operator had decided to add the Stars and Stripes — absent in the original image — in order to convey a meaning to the Keypic (Figure 4).

3 Spanning Concepts

A completely different idea for implicit concept description assigns the responsability of filling the semantic gap to the querying person. This was done in the *Trittico* experiment [2] (and is presently evolving in the relevance feedback paradigm [10]). The query is expressed by means of three images that the user either uploads or draws directly. The user is requested to propose three “extreme” instances of the shape he/she has in mind.

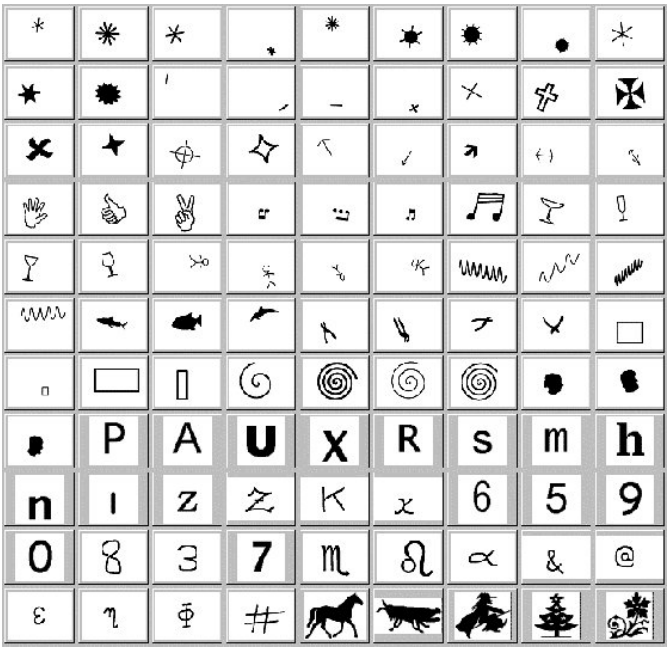


Fig. 5. A sample of the dataset

The mathematical core of the system is again a set of size functions (i.e. Persistent Betti Number Functions in degree zero). They depend on the choice of so called *filtering* (or *measuring*) functions (continuous real maps defined on the set). The definition of particular filtering functions allows to isolate those aspects of the image shape, which are of interest for the particular application goal.

Here is how the system should extrapolate the abstract shape concept from the three images. For each filtering function embedded in the program, the system computes the size functions of the three images, and evaluates their reciprocal distances. Then a comparison is done with the probability that these distances occur in a random triple. The lower the probability, the higher the weight that is given to the measuring function. The classifiers corresponding to the measuring functions, co-operate — with contributions depending on these weights — in the determination of a pertinence factor. On the base of the pertinence factor a set of images is extracted; finally, this set is sorted by a much finer comparison of the size functions.

In other words, the user puts a common feature into the three pictures, which for all the rest should be as different as possible. Then, those filtering functions which are best fit to recognize that feature, are stressed in the comparison of the three input images (or, better said: of their size functions) with the database.

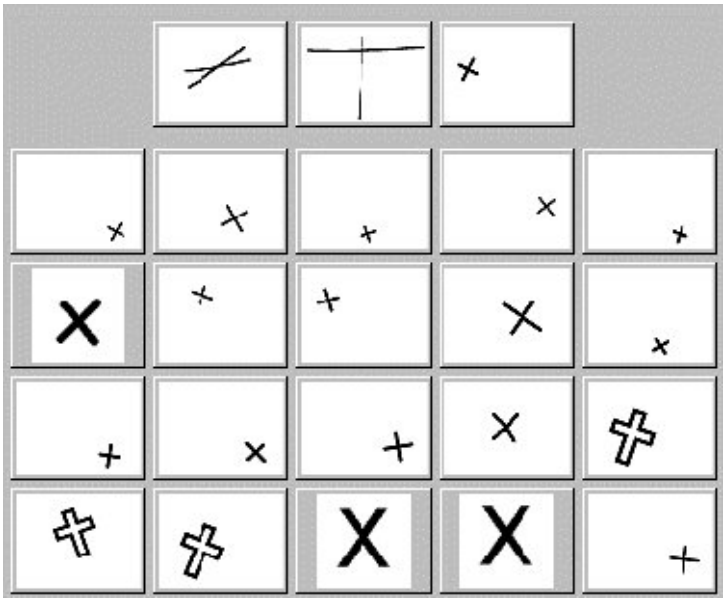


Fig. 6. A query and the first 20 output images

For checking this idea, we have chosen to build a database of simple black-on-white silhouettes. It consists of 2976 images belonging to 18 classes, plus 62 classes corresponding to alphabet characters, plus an extra "jamming" set of 197 unrelated images: see Figure 5 for a sample. Again, we are not interested here in reporting the quantitative assessment of the experiment, which you can find in [2]. What we want to stress, is that the system actually rather often output pertinent images which were very different from the query, "understanding" —so to say — the concept spanned by the three query images.

This is the case of Figure 6, for instance. The query consists of three heterogeneous crosses; this very independence has given the system the possibility to tune well the weights by which to stress some filtering functions against the other ones. So the output yields more crosses, some of which very different from the ones of the query.

A nice example is also given by Figure 7. In the first query, the three images are similar, but with three different orientations. The first ten output images have, coherently, no preferred orientation. In the second query, however, the three sketched little men are all upside down. The first ten output little men are also upside down. Without any need of directly imposing the system a restricted transformation group to be respected, the common attitude of the query images was sufficient to select the filtering functions which privileged this restriction.

This idea is being used in a series of experiments on relevance feedback, with more sophisticated, multidimensional filtering functions and by tuning distances. A first attempt is reported in [10].

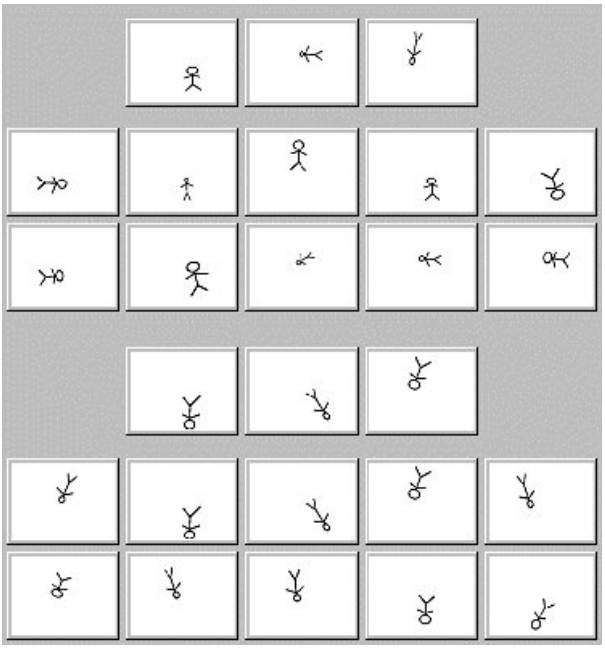


Fig. 7. Two queries and the first 10 output images

4 Conclusions

Two lines of research, in which human intervention bridges the semantic gap between image and concept, have been presented. In Keypics, image (and, more generally, document) annotation is by sketches rather than by words. Images are uploaded to the Web for a reason, and the image owner has all interest in focussing that reason, so he/she will extract from the images the part of content of his/her interest; this can be done by a simple, schematic picture drawn, copied or adapted by the image owner. In Trittico, a pictorial query is given by a triple of examples which span the searched for concept; the examples should be as far apart as possible, within the relevant class. Both lines have been developed by using persistent homology in degree zero (size functions), with successful initial experiments. The first awaits the sharing by a web users community for a true social experimentation. The second is evolving as a relevance feedback tool.

Acknowledgements. Work performed within the activity of INdAM-GNSAGA and of CIRAM and ARCES of the University of Bologna.

References

1. Biasotti, S., Cerri, A., Frosini, P., Giorgi, D., Landi, C.: Multidimensional size functions for shape comparison. Journal of Mathematical Imaging and Vision 32(2), 161–179 (2008)
2. Brucale, A., Cesari, F., d'Amico, M., Ferri, M., Frosini, P., Gualandri, L., Guerra, M., Lovato, A., Pace, I.: Image retrieval through abstract shape indication. In: Proceedings of the IAPR Workshop MVA 2000, Tokyo, November 28-30, pp. 367–370 (2000)
3. Cagliari, F., Di Fabio, B., Ferri, M.: One-dimensional reduction of multidimensional persistent homology. Proceedings of the American Mathematical Society 138(8), 3003–3017 (2010)
4. Cerri, A., Ferri, M., Frosini, P., Giorgi, D.: Keypics: free–hand drawn iconic keywords. International Journal of Shape Modeling 13(02), 125–137 (2007)
5. Cerri, A., Ferri, M., Giorgi, D.: A complete keypics experiment with size functions. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 357–366. Springer, Heidelberg (2005)
6. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inf. Syst. 29(2), 8:1–8:34 (2011), http://doi.acm.org/10.1145/1961209.1961211
7. Frosini, P.: Measuring shapes by size functions. In: Intelligent Robots and Computer Vision X: Algorithms and Techniques, Boston, November 01, pp. 122–133 (1991); International Society for Optics and Photonics (1992)
8. Frosini, P., Landi, C.: Size theory as a topological tool for computer vision. Pattern Recognition and Image Analysis 9(4), 596–603 (1999)
9. Frosini, P., Landi, C.: Persistent Betti numbers for a noise tolerant shape-based approach to image retrieval. Pattern Recognition Letters, 863–872 (2012)
10. Giorgi, D., Frosini, P., Spagnuolo, M., Falcidieno, B.: 3D relevance feedback via multilevel relevance judgements. The Visual Computer 26(10), 1321–1338 (2010)
11. Haav, H.M., Lubi, T.L.: A survey of concept-based information retrieval tools on the web. In: Proceedings of the 5th East-European Conference ADBIS, vol. 2, pp. 29–41 (2001)
12. Holzinger, A.: On knowledge discovery and interactive intelligent visualization of biomedical data. In: Proceedings of the Int. Conf. on Data Technologies and Applications DATA 2012, Rome, Italy, pp. 5–16 (2012)
13. Nake, F., Grabowski, S.: Human–computer interaction viewed as pseudo-communication. Knowledge-Based Systems 14(8), 441–447 (2001)
14. Norman, D.A.: Turn signals are the facial expressions of automobiles. Basic Books (1992)
15. Nowak, S., Hanbury, A., Deselaers, T.: Object and concept recognition for image retrieval. In: ImageCLEF, pp. 199–219. Springer (2010)
16. Obeid, M., Jedynak, B., Daoudi, M.: Image indexing & retrieval using intermediate features. In: Proceedings of the Ninth ACM International Conference on Multimedia, pp. 531–533. ACM (2001)
17. Reiterer, E., Dreher, H., Gütl, C.: Automatic concept retrieval with Rubrico. In: Multikonferenz Wirtschaftsinformatik, pp. 3–14 (2010)
18. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
19. Ziefle, M., Himmel, S., Holzinger, A.: How usage context shapes evaluation and adoption criteria in different technologies. In: AHFE 2012, Proceeding of Int. Conf. on Applied Human Factors and Ergonomics, San Francisco, pp. 2812–2821 (2012)

On Interactive Data Visualization of Physiological Low-Cost-Sensor Data with Focus on Mental Stress

Andreas Holzinger[1], Manuel Bruschi[1], and Wolfgang Eder[2]

[1] Medical University Graz, A-8036 Graz, Austria
Institute for Medical Informatics, Statistics & Documentation,
Research Unit Human-Computer Interaction
{a.holzinger,m.bruschi}@hci4all.at
[2] Wolfgang Eder Unternehmensentwicklung, A-8010 Graz, Austria
wolfgang.eder@eder.ch

Abstract. Emotions are important mental and physiological states influencing perception and cognition and have been a topic of interest in Human-Computer Interaction (HCI) for some time. Popular examples include stress detection or affective computing. The use of emotional effects for various applications in decision support systems is of increasing interest. Emotional and affective states represent very personal data and could be used for burn-out prevention. In this paper we report on first results and experiences of our EMOMES project, where the goal was to design and develop an end-user centered mobile software for interactive visualization of physiological data. Our solution was a star-plot visualization, which has been tested with data from N=50 managers (aged 25-55) taken during a burn-out prevention seminar. The results demonstrate that the leading psychologist could obtain insight into the data appropriately, thereby providing support in the prevention of stress and burnout syndromes.

Keywords: Data visualization, Knowledge Discovery, EDA, BVP, HRV, Stress, low-cost sensor.

1 Introduction

The value of emotion to the quality and range of everyday human experience is underestimated. It has a huge influence on the domains of cognition, in particular attention, memory, and reasoning [1]. An increasing problem in our western industrialized world is the Burnout Syndrome (BOS), which is a psychological state resulting from prolonged exposure to job stressors [2, 3]. In the past, many methods have been developed for measuring emotions, for a rapid overview refer e.g. to [4–7] and for a very short overview see [8], here a very brief summary: Electro-dermal activity (EDA), aka Galvanic skin response (GSR), electro-dermal response (EDR), or skin conductance response (SCR) is basically the measuring of the electrical resistance of the skin and can be used as a

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 469–480, 2013.

sensitive index of the activity of the sympathetic nervous system. Another popular method to assess the psychophysiological activity is Heart-Rate Variability (HRV) (see details in Section 2), which can be derived from Electrocardiographic (ECG) or Blood Volume Pulse Data (BVP). Viewing these Biosignals on mobile devices is a current and increasing trend [9].

Research on stress recognition and classification with physiological signals has reached a good point (for an overview look at [10]. Nonetheless, because people are sensitive to this topic due the popularity about burnout, we must be careful with our affirmations. Today, algorithms are able to reach a high accuracy for modeling stress but we still can not fully trust them, there is always the danger of modeling artifacts. However, stress remains an important aspect of human health that we must learn to deal with. Until recognition and classification patterns become fully reliable, we will continue to retrieve any data available and provide information to the individual expert end users, enabling them to gain knowledge by interactive visualizations - intelligence remains the forte of the human brain [11]. Moreover, while people are often unable to clearly identify their own emotions; it can be iteratively learned.

We recognized that the participants involvement/acceptance of the evaluation process necessitated an clear and easy to understand visualization of the physiological data processing results, in order for them to learn something about their emotions or discuss them with an expert.

Therefore our central goal was to design and develop an interactive information visualization for our signal processing model, which 1) displays most of the important feature evaluations regarding stress, 2) displays the data in a way that it is easy to compare features between relaxing and activating situations; 3) displays the data in a way that can be helpful for a non-expert end user and 4) at least may be provided with average computer resources within a tolerable time.

In short, the visualization must provide a variety of information about individual physiological activities with the focus on stress, so that both the non-expert and the expert can analyze and discuss the data and, most of all, to obtain insight and new knowledge from it. All the data collection, evaluation and visualization should be done by one user interface. To ensure that the solution will be affordable for many people, we have used a low-cost-sensor, which is described in Section 3.

2 Background - Psychophysiological Assessment

Physiological events are involuntary activities for which our Autonomic Nervous System (ANS) is responsible. Two main nervous systems are relevant for stress, the Sympathetic and Parasympathetic Nervous Systems (SNS and PNS respectively). Stressful situations cause dynamic changes in the ANS whereby the activity of the SNS increases and of the PNS decreases. In short, the SNS dominates during restless activities and the PNS during resting ones. These two systems are important for our research, because they regulate the different

physiological signals, such as heart rate variability (HRV), galvanic skin response (GSR), brain activity (EEG), blood pressure (BP) etc. Note that these systems are influenced by many different factors; two of them are eustress and distress. Eustress characterizes positive states and distress negative states, therefore, not all monitored stress should be perceived as bad stress [12].

As Sharma and Gedeon showed in [10]-(Table 5) HRV and GSR are two good parameters for detecting stress. In our work we used GSR and BVP records from a low-cost sensor. The signal of the BVP can be used to compute the HRV and some other features explained in Section 1.

2.1 Electrodermal Activity (EDA)

Electrodermal phenomena and the cardiac response are the most frequently assessed indices for the highly complex autonomic nervous system (ANS) activation in psychophysiology [13, 14]. The ease of obtaining a distinct electrodermal response (EDR) with inexpensive methods, the non-intrusiveness, and lenient field conditions are the major reasons for its popularity. Electrodermal recordings can use either external current (AC or DC) or the body's own electric current, the first are called exosomatic and second endosomatic. When an external current is applied to biological tissues such as skin, they act similar to electrical networks built of resistors and capacitors [13].

The term electrodermal activity (EDA) stands for all electrical phenomena of the skin, and was first introduced by Johnson and Lubin [15]. But there are also other frequently used terms, such as like Galvanic Skin Response (GSR). This electrical phenomena also includes all active and passive electrical properties which correlate to the skin and its extremities. The EDA has a central importance in biosignal acquisition and this concerns its psychological significance. Since the first research activities this response system has been closely linked with the psychological concepts of emotion, arousal and attention [14]. Basically it measures the hydration in the epidermis and dermis of the skin, which increases or decreases with the activation or inhibition of the sweat glands that are controlled by the sympathetic chain of the ANS.

Typically, this is recorded using two sensors placed at the surface of the hand or feet, since these are the areas of the body with higher seat gland density. In most cases the ring and middle finger are chosen and the most common unit used is μS.

The signal is a good indicator for stress, so that it helps to differentiate between conflict and non-conflict situations, but it is also a good indicator for dynamic activity. Usually a rapid rise of skin conductivity reflects a simple stress stimulus.

2.2 Cardiac Activity

The main purpose of our cardiac activity is to maintain our organs activity by providing them with blood, which is pumped around the body by the heart. When we are under stress, the heart rate is increased by the SNS and after

the stress has passed the PNS decreases it. It is therefore obvious that we can evaluate stress by measuring the rate of cardiac activity [16]. Regarding stress, as in [17], acute stress causes the heart to contract with high force and increased frequency. It is also known that with more chronic stress, the mass of the heart is increased. However, the baseline cardiac activity depends on the fitness of an individual and his activity.

There are several methods of measuring the cardiac activity, one used in our work is blood volume pulse (BVP). Photoplethysmography (PPG) is a non-invasive monitoring technique that can be used to track changes in the cardiac system. A reflective finger PPG sensor converts the fluctuation in the blood volume within a region of the index finger into a continuous waveform known as the Blood Volume Pulse (BVP). Traditionally, the BVP period was used to determine the heart rate. Current research however, shows that the BVP is capable of reflecting more than just the heart rate [18]. With the BVP signal, we are able to compute the RR-Intervals (the time between two following heart beats) [19] and therefore the Heart Rate Variability (HRV). More information is provided in Section 4.

3 Experimental Setup

Fifty healthy managers between the ages of 25 - 55 were recruited during a seminar and participated in our tests. Due to the pressure of time caused by testing this many participants during a seminar we had to design an experimental setup with a maximum time window of ten minutes. Therefore, the setup and introduction had to be fast and easy but without putting the participants under stress. Since we measured EDA and BVP, other requirements to our design arose. In order to have truthful signals, we had to ensure that there were no distracting or exciting elements in the room and also the temperature had to be kept at a pleasant level. In the past we experienced, some participants becoming nervous only because of a little blinking led on the notebook or the integrated webcam and sometimes the room temperature distorted our measurements.

As showed in [20] electrodermal and cardiac activity are both influenced by the physical activity of the participant. Both are strongly affected by anxiety and exercise [21] and in order to differentiate mental stress from other elicitations the activity also has to be considered. There are several solutions for this problem: firstly to add an accelerometer to the design, secondly to differentiate stress through feature extraction and preferably by reducing the physical activity as much as possible.

Another challenge we had set ourselves was to use low-cost sensors. Since the main purpose of our measurement is not to make a clinical analysis of the participants, but to visualize important information about their inner processes for discussion and learning, it is possible to keep the costs low.

Keeping all this in mind, we designed a hardware/software methodology that satisfies all these requirements.

At the time we designed the system according to the requirements, we chose the IOM Device from Wild Divine [22] for the acquisition of physiological signals.

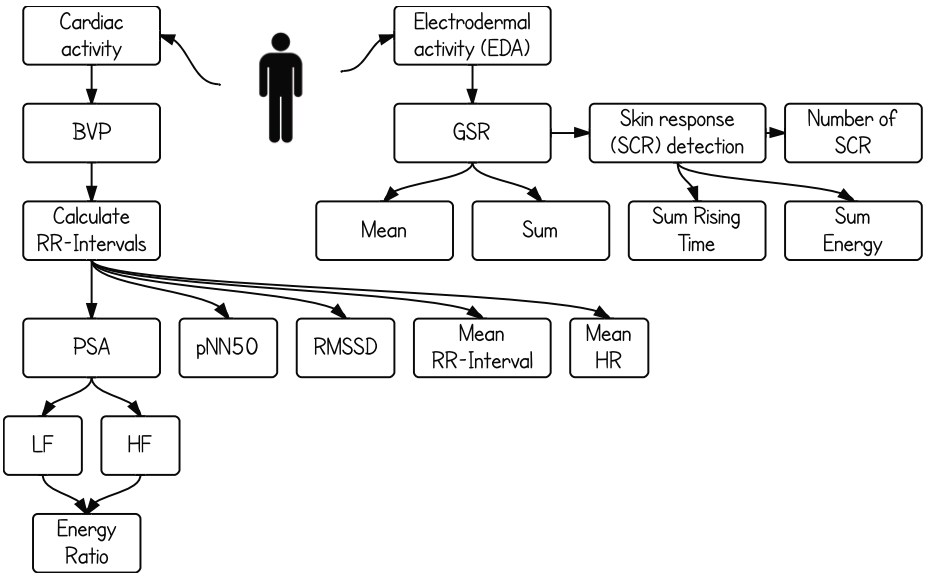


Fig. 1. Overview of extracted signal features

The device has been connected to a notebook on which we run our software. This shows a video file, logs the raw data from the sensor and write it to a file. With a video consisting of three parts, we were able to reduce the physical activity to a minimum because the participants were only asked only to observe and move as little as possible. The first part showed a slowly moving picture of a sky so that the subject could relax, after a while a calm voice announced the second part and mentioned that it will be stressful, the second part showed some random stressful pictures and colors with noisy sounds, after 15 seconds the voice soothed the participants and the third part with a slowly moving grass video, followed. During all three parts, relaxing music played in the background.

The chronological order of the test session was as following: During the first part (Introduction and Setup - 2 min.), the participants were made familiar with the system and the tasks. We also gave them some information about the three parts shown in the video. Then we asked them to sit comfortably, keep their hands still and watch the video. To avoid measurement falsification we left the room and reentered after the end of the video. This second part (Videotest) lasted 5 min. and the third and last part was the feedback part (3 min.).

4 Feature Extraction

In accordance with to many research reports about this topic [20, 23, 24], we chose to extract from our signals the features shown in figure 1. With a BVP signal, it is possible to calculate the RR-Intervals and from that we can compute the other features such as:

- Mean RR-Interval and mean HR: Each of them varies under stress and therefore reflects sympathetic or parasympathetic activities. Mean RR is significantly lower during a mental task than in the control condition [25]. A higher RR-Interval means a lower HR and vice versa. A significant change in the RR-Interval during stressful situations reflects a high HRV and therefore how well individuals are able to adapt to changes [26].
- Power Spectrum Analysis: As in [27], the high frequency (HF) is thought to reflect parasympathetic tone, whereas the very-low-frequency (VLF) and low-frequency (LF) are thought to reflect a mixture of parasympathetic and sympathetic tone. Because VLF has been found to distort stress detection [16], we left this out. With LF and HF we can also compute the energy ratio (total LF over total HF), which increases if stress levels increase [28].
- pNN50 and RMSSD: Both are time-domain related features that reflect parasympathetic activity [29, 30]. Since pNN50 is significantly lower with a mental task than in a control condition, it reflects mental stress [25].

GSR is directly influenced by the ANS and therefore it is overall a good indicator for stress. Already, minimal calculations, such as mean and sum are strong features. The detection of the skin responses (SCR), in order to compute the other three features shown in figure 1-(under SCR), is more complex but provides us with further indicators of stress [12, 28].

In total, we extracted twelve features and with this paper we suggest an easy to understand information visualization with the advantage of having a clear but deep insight into the data, in order to gain new knowledge.

5 Data Visualization

5.1 Related Work

We have had experience with using Star Plot diagrams, which are suitable for mobile and touch computers [31]. Star Plots aka radar charts [32], also called spider web diagram, polygon plot, polar chart, or Kiviat diagrams [33], are graphical methods of displaying multivariate data in the form of a 2D chart of three or more quantitative variables represented on axes starting from the same point. Each multivariate observation can be seen as a data point in an n-dimensional vector space:

- Arrange N axes on a circle in $\mathbb{R}^2$
- $3 \leq N \leq N_{max}$
- Map coordinate vectors $P \in \mathbb{R}^2$ from $\mathbb{R}^N \rightarrow \mathbb{R}^2$
- $P = \{p_1, p_2, ..., p_N\} \in \mathbb{R}^N$ where each p_i represents a different attribute with a different physical unit
- Each axis represents one attribute of data
- Each data record, or data point P is visualized by a line along the data points
- A line is perceived better than just points on the axes

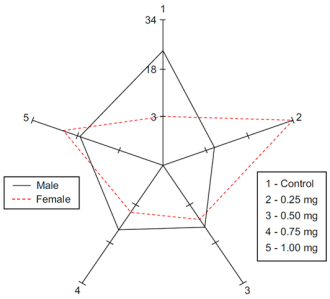


Fig. 2. A typical starplot diagram [34]

5.2 Our Solution

During the measurements, the participants passed through three phases: First a relaxing, second a stressful and finally a relaxing/recovering phase. Consequently, our idea was to use this data and visualize the results corresponding to these three parts using three different colors. The idea becomes clear in figure 3. Unfortunately, it can be seen only in grey scales.

The upper half of figure 3 refers to the cardiac signal and the lower half to the electrodermal activity. Whereas the three smaller Star Plot diagrams show the feature values of the corresponding phase, the bigger spider diagrams with the darker areas on the right show the overall feature values. The values of each feature are only displayed if a user hovers a feature caption. As an example, here in the figure it is the mean HR. If a user hovers one caption, the corresponding value is shown in all four diagrams. If a user makes a right-click on a caption, they receive a short textual introduction to the meaning of this feature regarding stress.

Both the feature values of the four Star Plots of the cardiac signal and the electrodermal activity part are scaled the same way. The upper limit corresponds to the max overall value and the lower limit to the minimal overall value. This allows us to better compare the Star Plots.

6 Results and Discussion

In the EMOMES project, we have tested 50 participants so far and provided feedback to them after the test, only by showing them two diagrams: one for the heart-rate and one for the galvanic skin response. We noticed that most of them were very interested in such insights and, as far as they could understand the theoretical concepts behind it within the short time frame given, some have provided supporting information for the signal interpretation and confirmed our hypothesis of the need of individual preliminary knowledge for the interpretation of the measured data. Just an interesting example: on one occasion, we noticed

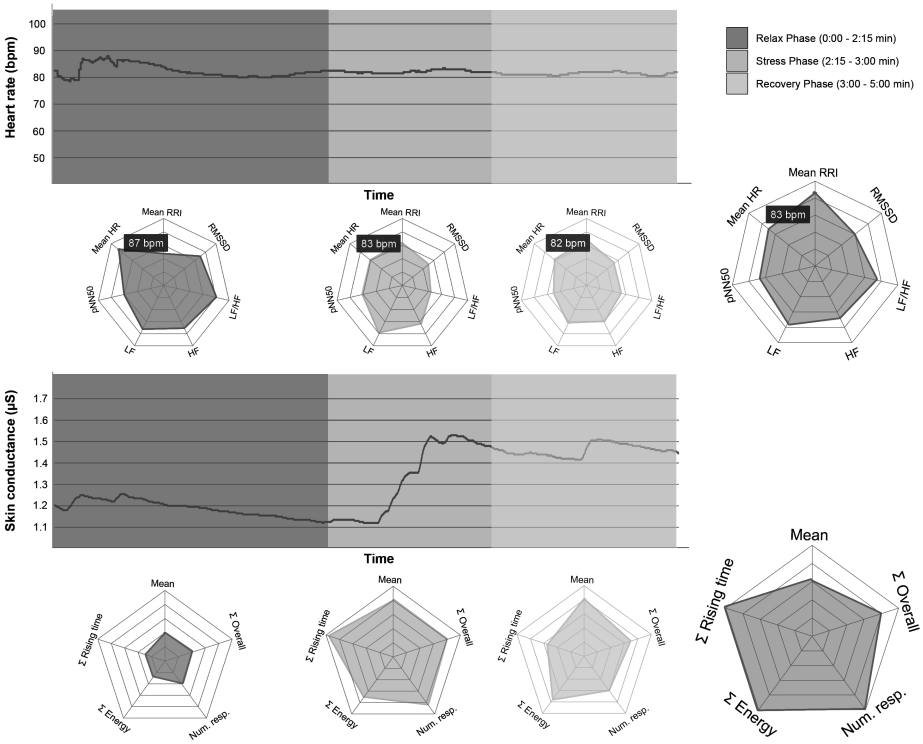


Fig. 3. Example of the information visualization screen

a high skin conductance response, which at first sight seemed similar to a measurement error, however, the participant told us that at that moment the music sounded like a mosquito and because he had a fight with one during the previous night, he got very nervous and flicked his finger.

In support of our hypothesis we extracted 12 features from the signals collected from these 50 participants. The signals were processed twice: once for the overall test and once for the three different parts. Therefore we divided the signals into three data entries each. With this data we provided the information needed by our visualization. The validity of the feature extraction is not given at this time as the low signal quality makes further examination necessary before all results can be confirmed, and only examples are offered. Nonetheless, because our focus is actually the visualization after the feature extraction, we had the necessary data structure for the visualization given in figure 3.

Results from our signal processing model are shown in table 1 and are from all three video parts together. Actually, it is not possible to measure in this short time frame if someone has a BOS and therefore we can not claim that this visualization tells whether someone is afflicted by BOS or not. However, providing this clear overview to experts gives them a good objective insight

Table 1. Results from our Signal Processing Model

Feature	**Min**	**Mean**	**Max**
MeanRR (ms)	669	867	1114
MeanHR (bpm)	54	73	93
pNN50 (%)	24.80	49.44	76.90
RMSSD (ms)	46	111	265
LF (%)	23.90	40.69	56.70
HF (%)	16.50	39.13	68.30
LF/HF	0.364	1.2480	3.056
Sum GSR (μS)	-3.88	0.4640	3.09

about individual stress reactions and has been confirmed as a good starting point to discuss the autonomous stress behaviours and preventing stress and burnout syndromes in order to decide whether further investigations are necessary or not.

As mentioned in section 4 we can tell something about stress activity by looking at these features. As an example, if you look at the table 1 the sum of GSR tells us whether or not the participant was able to recover from the stress within this time frame. If this value is positive, it means that he had more SNS activity than PNS and vice versa. The same is for the LF/HF-Ratio. If it is high, it means the participant had a low HF activity and therefore less parasympathetic activity. By splitting up the feature extraction in three parts in correlation to the three video parts, we are able to compare them in more detail and have a deeper insight into the physiological activity. This provides us with the possibility to give more detailed feedback to the participants and a more profound basis for discussions and analysis.

7 Conclusion and Future Research

In this work, we described our lessons learned from testing 50 participants and we present a visualisation for providing insight into physiological behaviours regarding stress.

These goals were achieved to our satisfaction. The visualization displays most of the important feature evaluations regarding stress and displays them in a way that it is easy to compare features between the three parts and the overall features. It gives us a very clean insight into the data and offers good possibilities for knowledge discovery. With a minimum of support, a layman is able to understand the science behind it and can support us with useful information. Nonetheless there are some things that can be improved and expanded.

The low-cost sensor provided workable signals, but for more accurate studies the signal quality is too low. However there are still low-cost sensors that are

promising, such the BITalino [35] and for better results for the visualization others will also be tried. After the completion of the processing, the evaluation of this visualization with a larger test group will provide further insight into our suppositions.

For a better insight for stress prevention, it would be helpful to include in our software a questionnaire like the Hamburger Burnout Inventory (HBI) that reflects the subjective sensation of the participants, especially looking at BOS. If the participants do this in advance of the test, at the end they have an objective and subjective insight.

Acknowledgments. We thank the anonymous reviewers for their helpful comments. Many thanks also to Hugo Silva from the Technical University of Lisbon for constantly supporting our research. Thanks to Carl Bck from the Technical University of Graz for some helpful comments. This work was partially funded by Austrian Science Fund (FFG), Innovationsscheck Plus No. 840131.

References

1. Dolan, R.J.: Emotion, cognition, and behavior. Science 298(5596), 1191–1194 (2002)
2. Le Gall, J., Azoulay, E., Embriaco, N., Poncet, M., Pochard, F.: Burn out syndrome among critical care workers]. Bulletin de l'Académie Nationale de Médecine 195(2), 389 (2011)
3. Weber, A., Jaekel-Reinhard, A.: Burnout syndrome: a disease of modern societies? Occupational Medicine 50(7), 512–517 (2000)
4. Wickens, C.D., Gordon, S.E., Liu, Y.: An introduction to human factors engineering (2004)
5. Rubin, J., Chisnell, D.: Handbook of Usability Testing: Howto Plan, Design, and Conduct Effective Tests. Wiley (2008)
6. Cairns, P., Cox, A.L.: Research methods for human-computer interaction. Cambridge University Press (2008)
7. Lazar, J., Feng, J.H., Hochheiser, H.: Research methods in human-computer interaction. Wiley (2010)
8. Holzinger, A.: Process Guide for Students for Interdisciplinary Work in Computer Science/Informatics: Instructions Manual-Handbuch für Studierende. BoD–Books on Demand (2010)
9. Breitwieser, C., Terbu, O., Holzinger, A., Brunner, C., Lindstaedt, S., Müller-Putz, G.R.: iScope – viewing biosignals on mobile devices. In: Zu, Q., Hu, B., Elçi, A. (eds.) ICPCA 2012 and SWS 2012. LNCS, vol. 7719, pp. 50–56. Springer, Heidelberg (2013)
10. Sharma, N., Gedeon, T.: Objective measures, sensors and computational techniques for stress recognition and classification: A survey. Computer Methods and Programs in Biomedicine 108(3), 1287–1301 (2012)
11. Holzinger, A.: On knowledge discovery and interactive intelligent visualization of biomedical data-challenges in human-computer interaction & biomedical informatics. In: 9th International Joint Conference on e-Business and Telecommunications, ICETE 2012, pp. IS9–IS20 (2012)

12. Boucsein, W.: Electrodermal activity. Springer (2012)
13. Boucsein, W.: Electrodermal Activity. Springer (2012)
14. Dawson, M.E., Schell, A.M., Filion, D.L.: The electrodermal system. In: Cacioppo, J.T., Tassinary, L.G. (eds.) Handbook of Psychophysiology, 3rd edn., pp. 159–181. Cambridge Press (2007)
15. Johnson, L.C., Lubin, A.: Spontaneous electrodermal activity during waking and sleeping. Psychophysiology 3(1) (1966)
16. Camm, A.J., Malik, M., Bigger, J., Breithardt, G., Cerutti, S., Cohen, R., Coumel, P., Fallen, E., Kennedy, H., Kleiger, R., et al.: Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. Circulation 93(5), 1043–1065 (1996)
17. Devereux, R.B., Roman, M.J., Palmieri, V., Okin, P.M., Boman, K., Gerdts, E., Nieminen, M.S., Papademetriou, V., Wachtell, K., Dahlöf, B.: Left ventricular wall stresses and wall stress-mass-heart rate products in hypertensive patients with electrocardiographic left ventricular hypertrophy: The life study. Journal of Hypertension 18(8), 1129–1138 (2000)
18. Barreto, A.B., Aguilar, C.D., Jakubzick, E.E.: Adaptive lms delay measurement in dual blood volume pulse signals for non-invasive monitoring [photoplethysmography]. In: Proceedings of the 1997 Sixteenth Southern Biomedical Engineering Conference, pp. 117–120. IEEE (1997)
19. Silva, H., Sousa, J., Gamboa, H.: Study and evaluation of palmar blood volume pulse for heart rate monitoring in a multimodal framework. Computing Paradigms for Mental Health, 35 (2012)
20. Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., Griss, M.: Activity-aware mental stress detection using physiological sensors. In: Gris, M., Yang, G. (eds.) MobiCASE 2010. LNICST, vol. 76, pp. 282–301. Springer, Heidelberg (2012)
21. Wilhelm, F.H., Pfaltz, M.C., Grossman, P., Roth, W.T.: Distinguishing emotional from physical activation in ambulatory psychophysiological monitoring. Biomedical Sciences Instrumentation 42, 458–463 (2006)
22. Wilddivine (2013), `http://www.wilddivine.com` (accessed May 01, 2013)
23. Zhai, J., Barreto, A.: Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In: 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2006, pp. 1355–1358. IEEE (2006)
24. Canento, F., Silva, H., Fred, A.: Applicability of multi-modal electrophysiological data acquisition and processing to emotion recognition. Computing Paradigms for Mental Health, 59 (2012)
25. Taelman, J., Vandeput, S., Spaepen, A., Van Huffel, S.: Influence of mental stress on heart rate and heart rate variability. In: 4th European Conference of the International Federation for Medical and Biological Engineering, pp. 1366–1369. Springer (2009)
26. Acharya, U.R., Joseph, K.P., Kannathal, N., Lim, C.M., Suri, J.S.: Heart rate variability: a review. Medical and Biological Engineering and Computing 44(12), 1031–1051 (2006)
27. Lauer, M.S.: Autonomic function and prognosis. Cleveland Clinic Journal of Medicine 76(suppl 2), S18–S22 (2009)
28. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems 6(2), 156–166 (2005)

29. Buchheit, M., Papelier, Y., Laursen, P.B., Ahmaidi, S.: Noninvasive assessment of cardiac parasympathetic function: postexercise heart rate recovery or heart rate variability? American Journal of Physiology-Heart and Circulatory Physiology 293(1), H8–H10 (2007)
30. Zulfiqar, U., Jurivich, D.A., Gao, W., Singer, D.H.: Relation of high heart rate variability to healthy longevity. The American Journal of Cardiology 105(8), 1181–1185 (2010)
31. Holzinger, A., Höller, M., Bloice, M., Urlesberger, B.: Typical problems with developing mobile applications for health care. In: ICE-B 2008, p. 235 (2008)
32. Ebert, P.S.: Smart radar chart. EP Patent 1,530,141 (May 11, 2005)
33. Morris, M.F.: Kiviat graphs: conventions and figures of merit. ACM SIGMETRICS Performance Evaluation Review 3(3), 2–8 (1974)
34. Saary, M.J.: Radar plots: a useful way for presenting multivariate health care data. Journal of Clinical Epidemiology 61(4), 311–317 (2008)
35. Alves, A., Silva, H., Lourenço, A., Fred, A.: BITalino: A Biosignal Acquisition System based on the Arduino. In: Proceeding of the 6th Conference on Bio-Inspired Systems and Signal Processing, BIOSIGNALS (2013)

Marking Menus for Eyes-Free Interaction Using Smart Phones and Tablets

Jens Bauer[1], Achim Ebert[1], Oliver Kreylos[2], and Bernd Hamann[2]

[1] Computer Graphics and HCI Lab, TU Kaiserslautern,
67663 Kaiserslautern, Germany
j_bauer@cs.uni-kl.de
[2] Institute for Data Analysis and Visualization, Department of Computer Science,
University of California, Davis, CA 95616, U.S.A.

Abstract. Large displays are helpful tools for knowledge discovery applications. The increased screen real estate allows for more data to be shown at once. In some cases using virtual reality visualizations helps in creating more useful visualizations. In such settings, traditional input devices are not well-suited. They also do not scale well to multiple users, effectively limiting collaborative knowledge discovery scenarios. Smart phones and tablet computers are becoming increasingly ubiquitous and powerful, even having multi-core CPUs and dedicated Graphic Processing Units (GPUs). Given their built-in sensors they can serve as replacements for currently-used input devices, and provide novel functionality not achieved with traditional devices. Furthermore, their ubiquity ensures that they scale well to multi-user environments, where users can use their own devices. We present an application-independent way to integrate smart phones and tablets into knowledge discovery applications as input devices with additional functionality. This approach is based on Marking Menus, but extends the basic idea by employing the special capabilities of current consumer-level smart phones and tablets.

Keywords: input devices, collaborative interaction, multimodal interaction, 3D Interaction.

1 Introduction

Smart phones and tablet computers are becoming increasingly popular and powerful. Current consumer-level devices feature multi-core CPUs and dedicated GPUs. They contain multi-touch screens, GPS receivers, compasses, and accelerometers, and offer connectivity via WiFi, bluetooth, and 3G technologies. Given those capabilities, smart phones and tablets are attractive alternatives to "traditional" input devices. Smart phones and tablets offer several benefits, due to their ubiquity and added functionality compared to typical input devices.

As these devices are multi-functional and wide-spread, most users of knowledge discovery applications already own at least one, which precludes the need to buy often expensive single-purpose input devices. The same holds for multi-user

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 481–494, 2013.

environments, where each user can use his/her own device to control the system. Due to the devices' portability, they could potentially be used as (limited) "mobile memory" to transfer data between environments or as a kind of token to identify users.

Non-trivial applications typically provide the user with a moderate to large number of functions, which need to be mapped to the set of available input devices. Normally, each function is mapped to one device button, or gesture, or user interface (UI) button displayed. Problems arise when the number of required functions exceeds the number of input device buttons or gestures, or when the number of displayed UI buttons clutters the display. The touch-capable screens of current mobile devices, on the other hand, provide enough screen real estate to offer a large number of buttons and mapped functions – significantly more than practical for traditional input devices. However, the naïve approach of using mobile devices disrupts users' workflows, as they have to shift their attention back and forth between the devices' small screens and the main display environment.

We propose to utilize *Marking Menus* [16], a radial menu structure, as a central element of a novel interaction method employing the multi-touch screen of mobile devices. The touch screen is used to show (hierarchical) radial menus as they pop up. This enables eyes-free interaction for experienced users who do not need the visual feedback from the mobile device, and leads to increased efficiency for those users, while at the same time keeping the menu structures visible should they be needed. This selection method is extended by the usage of tracking sensors, accelerometers, multi-touch and/or in-menu slider controls to provide a larger variety of interaction possibilities, which are explained in detail later in this paper.

The advantages of this design over existing interaction methods are:

- Eyes-free interaction makes complex user interaction possible without interrupting the workflow.
- Auditory or haptic feedback is given to support eyes-free interaction even more.
- The general approach allows application of the design to a wide array of new and existing applications.
- The ubiquity of smart phones and tablets and their usability for other tasks make this approach very cost-effective.
- As the system uses its own mobile screen it can replace large parts of the application's graphical user interface (GUI) up to the complete GUI in some cases.
- Support for multiple users and the system scales well to the number of users.

We present the design and implementation of our prototype. A formal user study is beyond the scope of this paper, but is planned as part of our future work. Studies by Kurtenbach et al. [17] are applicable to this interaction method and show its general usability.

2 Related Work

2.1 Remote Interaction Using Mobile Phones

The use of mobile or smart phones as input device has been an active area of research. Ballagas et al. [3] presented two interaction techniques for camera-equipped phones. One uses the phone as a replacement for an optical mouse, with the physical x-y-translation of the phone mapped to a traditional cursor on the screen. At the time, the method incurred a latency of about 200 ms, too much for a practical applications; even today, high latency seems to be an issue for such approaches. This approach also does not increase the number of functions that can be mapped to a device. The second technique in the same paper is a pure selection method, where the phone detects at which object on the screen it is pointed by tracking markers displayed on the screen. These markers only appear for brief moments while the phone's camera is active to reduce display clutter. Both approaches only employ camera phones as direct mouse replacements, missing the opportunity to further improve the available interaction.

Extending that work, Jeon et al. [13] proposed interaction techniques to replace the computer mouse with mobile phones. The movement of the phone in the air is mapped either to a cursor, or directly to an interactive object. Three approaches of calculating the relative movement of the phone are described: motion-flow, marker tracking with the marker on an selectable object, and marker tracking with a free-floating marker. While this is more refined than Ballagas' et al.'s approach, it still suffers from the same basic limitations.

A specialized approach is presented by Madhavapeddy et al. [19]. A camera-equipped phone is used to control a flight booking application, where the starting and destination airports are selected by pointing at them with the phone, again using on-screen tracking markers. A similar approach was used by Thelen et al. [23] to interact with a 3D representation of the human brain. Here, 2D markers are rendered as part of a 3D model of the human brain, and scanning one of those markers with a smart phone displays additional information associated with that marker. Both are examples of usable designs, but have the drawback of being very specific to a certain kind of task, i. e., selection from a small number of preset targets.

2.2 Marking Menus

Kurtenbach et al. [16,17] proposed and evaluated *Marking Menus* as an improved version of radial menus. The main difference between Marking Menus and transitional radial menus is the absence of a completely bounded target area for each menu item in the former. Radial menus simply arrange their items in a circular pattern around a center point in one or more "rings." An item is selected when the selection cursor is within the bounds of the item, and selection is usually affirmed by a button press on the input device (or any other available confirmation gesture). If the selected menu item has subitems, this causes a new radial

menu to pop up at or near the current cursor position. Marking Menus only have a single ring of items, and their respective target areas expand infinitely outwards from the center of the menu in a wedge-like shape. Holding the cursor still in the selection area of an item with sub-items pops up another Marking Menu showing the sub-items, and a confirmation event (button press, or, in the original example, lifting the pen from the display surface) selects the current item. The main benefit of this menu layout and selection mechanism is eyes-free item selection, which was at the time proposed to address the high latency of pen-based direct interaction displays on then-current workstations. Using Marking Menus, users could either put the pen down on the screen, wait for the menu to appear, and select items and sub-items by drawing a stroke from the center into the (wedge-shaped) selection areas, while experienced users could just draw the chain of strokes used to reach a certain item without even waiting for the menu to pop up. (Kurtenbach called this "Expert Mode".) Since the selection areas are theoretically infinitely large, user accuracy is very high and Marking Menus are very easy to use. These assertions are backed up by Fitts' Law [5] and Steering Law [1]. While the original high-latency problem Marking Menus were designed to address no longer matters, their effectiveness still does, and Marking Menus have been incorporated into many modern applications.

Pook et al. [22] proposed *Control Menus*, a general improvement that also applies to Marking Menus: instead of using menus only for selections, the continued motion of an input device after an item has been selected is used to change a continuous value associated with that item. They use an example of a "zoom" menu item, where any continued input device motion after selecting that item directly alters the current zoom level. This method could also be applied to to related continuous values by using both display axes simultaneously.

Perlin's *QuikWrite* [21] is a stroke-based text entry method. It lays out the available character set along the edge of a rectangle (usually covering the entire screen), and divides them into eight zones (N, NE, E, etc.). A character is selected by first moving the input device from the rectangle's center into a zone, then optionally into a second zone, and finally back to the center. For example, the top-left (NW) character of the top-right (NE) zone is selected by first moving into the NE zone, then into the NW zone, and then back to the center. QuikWrite allows to enter entire phrases of text with a single, continuous stroke of the input device.

Other touch text input methods include *Cirrin* [20] and *T-Cube* [24]. The former selects characters by touching them with a stylus and allows multiple selections with one single stroke, while the latter is a radial menu where the characters are arranged in multiple circles, so the direction and the length of a stroke together are responsible for character selection. Among these methods, only Quikwriting can seamlessly be used in a Marking Menu. Holzinger et al.[11] propose to use a stylus to aid handwriting on touch surfaces. This could also be integrated into a Marking Menu.

Guimbretière and Winograd [9] presented *FlowMenus*, a menu system designed for pens on large touch-sensitive displays. FlowMenus work like normal

Marking Menus for the first selection step, but selection in submenus of any level is based on curved stroke gestures back to the submenu's center: sloped clockwise, sloped counter-clockwise, and straight, supporting only three items per submenu. The advantage of this method is that the pen always returns to the original selected position, and therefore allows direct interaction with an on-screen element at that position after an item has been selected. However, in Expert Mode, it can be difficult to return precisely to the menu center, and small deviations may add up to a point where users might accidentally select a wrong menu item. FlowMenus also allow text input directly chained into the menu using QuikWrite, and value selection similar to Control Menus. But since submenu selection strokes need to curve back to the submenu's center, values now need to be selected based on the length of the curved stroke, which is less intuitive. This also shows a subtle menu design limitation of FlowMenus: menu items for direct object interaction and for text input need to be on the center of the menu structure and thus on an even level of the submenu hierarchy, while value-selection items need to be on an odd level for the same reason.

More new menu designs based on Marking Menus emerged recently. Bailly and Lecolinet [2] allowed curved strokes for menu item selection, resulting in so-called "Flower Menus." Depending on the initial stroke direction, the curvature and the curve direction, i. e., clockwise or counter-clockwise, different items can be selected. While this allows a higher density of menu items per submenu level, the total time needed to draw the curved strokes is notably higher than drawing just straight lines. Using a similar, but unrelated, design to Marking Menus, *Wavelet Menus* [7,6] are also controlled by straight strokes. The menu is radial and the menu grows in radius whenever a submenu is opened and placed on. New submenus are positioned innermost. The main difference to Marking Menus is that the user has to release the touch after each level of the submenu hierarchy and start each new stroke in the center of the screen.

A first attempt to improve Marking Menus with multi-touch was presented by Lepinski and Grossman [18]. Their design uses chording gestures to invoke up to 31 different menus, based on which fingers of a hand are touching the screen. However, due to technical limitations, this approach requires the user to put down all five fingers first, and then lift a subset of fingers afterwards to invoke a menu. This is awkward, causes unnecessary delay, leads to selection mistakes if users lift fingers faster than the system can detect their gestures, and prohibits other simultaneous uses of multi-touch. In the near future these technical limitations might no longer apply to large touch sensitive surfaces, but consumer level smart phones and tablets will probably lack the feature of tracking non-touching fingers for still quite some time.

Kin and Hartmann [14] studied user experience with two-handed Marking Menus. They proposed either to split the menus into parts for each hand, and let the user select items either simultaneously or sequentially, potentially increasing efficiency. Their method requires clutching with the fingers to select items deeper in the menu hierarchy as one stroke has to be completed by lifting the finger

from the surface in order to access the next level of menu items, and does not allow multi-finger gestures.

Heidrich et al. [10] used an approach similar to Marking Menus to control a Smart Home. The gestures are performed on a table monitored by a Kinect and the menu is projected on the table. While the basic approach is the same as described in this paper, Heidrichet al.'s method does not allow multi-touch or value selection.

3 Design

Smart phones and tablets are not only usable as input devices, but also have output capabilities. When designing a user interface, one has to decide whether, and how, to use those capabilities. A device's screen, typically the primary output channel, can be used in a variety of roles: as a primary screen for a focus-and-context displays [4]; as a full additional screen, as in Air Display[1]; to clone the environment's main display, as in remote access applications such as Remote Desktop Protocol (RDP) or Virtual Network Computing (VNC); or even not at all, as in mouse control applications such as Remote Mouse[2]. One reason not to use a device's screen is that having to shift focus between multiple unrelated screens might interrupt a user's workflow.

Our approach uses touch screens for interaction feedback via radial menus similar to Marking Menus, described in Section 2.2. Kurtenbach 2.2 categorizes the usage modes of Marking Menus as Novice Mode (waiting for menus to pop up) and Expert Mode (completing interaction before or while a menu pops up). In the remainder of this paper, we refer to Expert Mode as eyes-free mode, to emphasize the fact that it does not require shifting attention away from an environment's main display.

3.1 Menu Design

A radial menu is presented to the user when they touch the screen (see Fig. 1). Its design is similar to the original Marking Menu. The menu is centered around the initial finger position, and menu items are selected as the finger enters their selection area. If the selected item has sub-items, a submenu pops up on selection (see Fig. 2). Unlike regular radial menus, the submenus have an additional wedge-shaped *dead zone*, i. e., a zone where no selection is made, along the line from the center of the current submenu through the center of its parent menu. This dead zone improves usability in eyes-free mode: when not looking at the screen while interacting, users run the risk of making shorter or longer finger movements than intended. To account for that, submenus initially move with the user's finger until the movement changes direction. This ensures that stroke length does not influence item selection.

[1] `http://www.avatron.com`
[2] `http://www.remotemouse.net/`

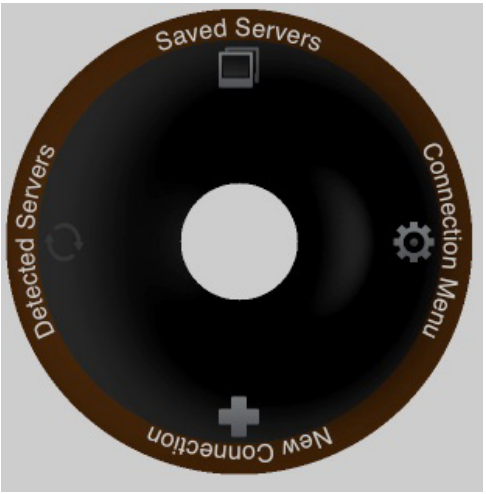


Fig. 1. The general design of the menu, in this example for four menu items, all with a descriptive name and an icon. .

The wedge-shaped dead zone has the additional benefit of allowing to undo selections, as users can backtrack their strokes through multiple levels of the submenu hierarchy, including canceling menu invocation entirely. This is an extension of the original Marking Menu method. To support eyes-free interaction, the phone can provide auditory or haptic feedback whenever the user tracks back one submenu level.

To avoid unintentional selections caused by touch screen jitter, another circular dead zone has to be defined in the center of the menu. This dead zone should be as small as possible, as it defines the minimal length of a stroke to be recognized. The dead zone additionally improves selection accuracy, as the length of a stroke determines the accuracy of measuring that stroke's angle. Due to the rather coarse resolution of current-generation touch screens, a very short stroke may only be a few pixels long, resulting in inaccuracy when detecting the angle of that stroke. Thus, the size of the central dead zone is a device- and user-dependent configuration parameter.

To reduce display clutter in deep submenu hierarchies, where submenus are drawn on top of their respective parents, only the currently active submenu is drawn fully opaque, while parent submenus are drawn with increasing levels of transparency, i. e., the root menu is most transparent. Drawing the entire menu hierarchy in this way helps users to see their current position in the menu hierarchy and the direction of the most recent stroke, should they lose their place during eyes-free interaction. It also provides valuable feedback for the multi-touch interaction described in Section 3.3.

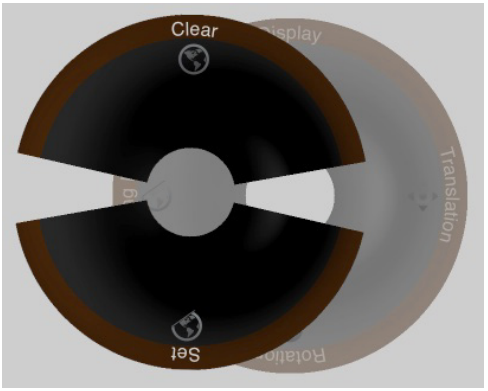


Fig. 2. The menu with a submenu opened up after the user moved their finger to the left. The submenu is on top of the half-transparent parent menu. Note the dead zone through in sub-menu to improve usability in eyes-free mode and to allow backtracking of selections.

3.2 Value Control

Our menu design also supports value selection similarly to Control Menus, as described in Section 2.2. Menu items with associated values are indicated by arrows (see Fig. 3). But instead of only allowing simple slider-like control over one or two dimensions, our design allows two modes of operation. In absolute mode it works like a regular slider control, reporting the absolute position of the touch(es) to the application. In relative mode, the actual location of the touch(es) are irrelevant. Instead only the changes in position are reported to the application. Absolute mode is useful for any fixed range of continues values, i. e., whenever a traditional slider control is useful. Relative mode is for situations where the absolute value of a variable is of less interest than the actual change to it. A common example for this is the position of an object. The user is normally more interested in moving the object a certain amount of units, instead of setting it to an absolute value. This relieves the application of keeping track of an absolute value (i. e., the position of the slider) that is of little interest.

For many visualization applications 1D or even 2D positioning is not enough. Employing the well-known multi-touch gestures of pinch and rotate, four degrees-of-freedom (DOF) can be controlled at once (x-Direction, y-Direction, Pinch and Rotation). This allows for simultaneous two-DOF-Movement, Zoom and one-DOF-Rotation, for example, to allow the user to drill down into a part of the visualized data without having to apply a new menu selection. Alternatively multiple quantifications can be made at the same time, with each finger touching the screen functioning as a separate one-DOF slider control, theoretically allowing for a 10-DOF control. The practical limit depends on actual touch-screen size, the size of the user's fingers and the user's dexterity. Most people

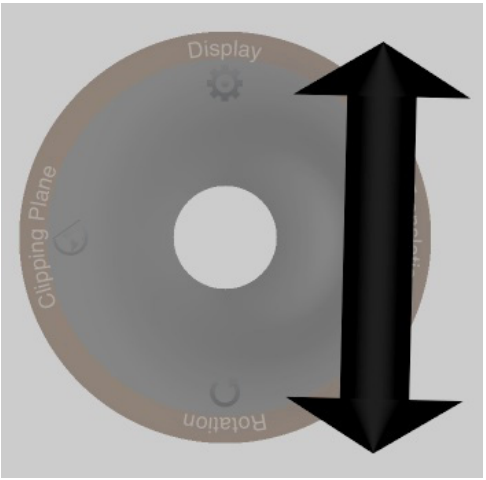


Fig. 3. An arrow shows up when the selected submenu supports direct value selection, just like a regular slider

should be able to use four sliders simultaneously without serious problems. The control provided might be too coarse or too fine. Therefore the control itself can be scaled using the pinch gesture with a second finger (one-DOF or two-DOF controls) or a third finger (three-DOF and four-DOF). The scaling is visualized by scaling the control on the touchscreen accordingly.

3.3 Multi-touch Capabilitiy

In our method, multi-touch is not only used for value control, but also for normal menu navigation. As shown in Fig. 4, putting down another finger on the touch screen while a menu item with sub-items is selected causes the submenu for that item to detach from its parent menu, to move to the new finger's position, and subsequently to be controlled by the new finger. The parent menu sticks to the original finger, and move with it, but is otherwise be locked as long as a detached submenu is active. If, on the other hand, the currently active menu item is a value selection item, then the new finger will invoke multi-DOF value selection as described in Section 3.2. If the currently selected item is a regular menu item, then each additional finger will cause a selection event for that menu item, enabling rapid multiple selection of the same menu item by repeated tapping with an additional finger.

Detachable submenus are useful when stroke navigation through a deep menu hierarchy reaches the edge of the touch screen at any point. In that case, the user can place a second finger on the opposite side of the touch screen, and continue with menu selection normally. To prevent awkward positions, the initial finger can be released once the second finger starts interacting with the menu hierarchy. If, however, the first finger is kept held down, then only the detached part of

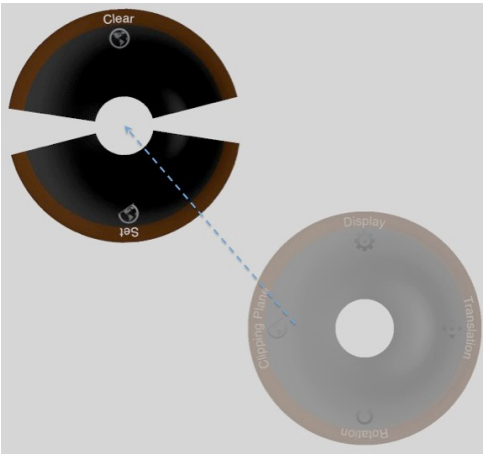


Fig. 4. Multi-Touch enabled navigation allows the user to detach a submenu and control it with another finger while the original menu still sticks to the old touch

the menu structure, and not the entire hierarchy, is dismissed after a successful selection. This enables a shortcut for multiple subsequent selections that have the same prefix in the menu hierarchy. For example, if the user wants to execute *left*, *up*, *left* followed by *left*, *up*, *right*, it will be possible to stroke *left*, *up* first, then perform *left* using a second finger, causing the menu at the first two parts of the stroke to be still open after the selection. With either the original or an additional touch the *right* stroke can be performed. A similar example is depicted in Fig. 5. Such shortcuts becomes more useful with deeper menu hierarchies. Very experienced "power users" can push the paradigm by using more than two fingers, detaching multiple submenus at the same time. With properly designed deep menu hierarchies, such expert shortcuts, while having a relatively steep learning curve, can lead to highly efficient interaction sequences.

The methods described up to this point support selection, quantification, an approximation of positioning and orientation via sequences of quantification, path via sequences of approximated position and orientation, and text when combined with QuikWrite or a comparable approach. True 3D interaction, however, requires at least 6 degrees of freedom, and single- or multi-touch gestures are not an intuitive replacement. While it is possible to chain two individual three-DOF interactions together to achieve six-DOF, this is difficult to handle for the user. Also it is only possible for two-handed interactions, and users should be able to choose one-handed vs. two-handed usage freely for themselves. Another way is to employ the smartphones build-in accelerometers to deliver the current orientation of the phone as additional information for each menu. As orientation provides three DOF (pitch, yaw and roll), seven DOF can be achieved through the combination of multi-touch quantification and orientation. Since 4 of these 7 DOFs concern rotation, a more practical limit is six DOF though. The

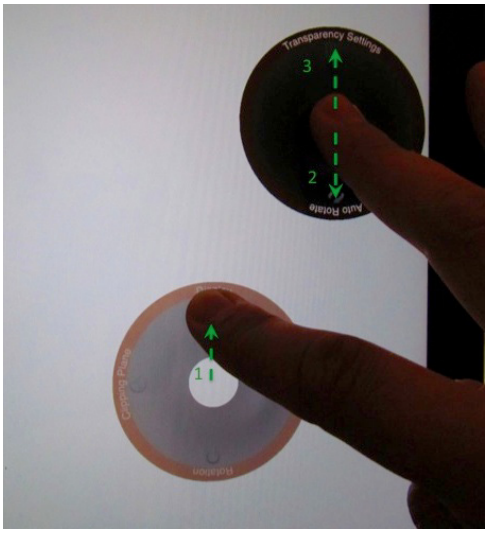


Fig. 5. Using detached menus as shortcut: The index finger touches the surface and is swept up (1). The middle finger swipes down causing the submenu to detach and toggle the auto-rotate function (2). The middle finger touches the surface another time, swiping up to activate the transparency settings (3).

quantification control already described can transmit the current orientation of the device as an additional 3 quantifications. This can be combined into free six-DOF-movement, for example, by letting the position of the user's touches control the x and y position, pinching of the fingers controls the z position (or zoom) and the device rotation can be transferred to the object being moved.

To support clutching (i.e., repeated swipes on the screen to control the same value) and to offer quick access to the last used function, the last invoked action can be part of the menu or can be activated by the use of multiple touches. For example a menu can be designed in a way that a single touch with a swipe to the left and then up lets the user control the position of some object. If the user is not satisfied with the object's position afterwards, a simple touch with two fingers will allow one to refine the positioning without having to swipe left and up again.

3.4 Hardware Requirements

Our design can be implemented on most current consumer-level smart phones and tablets. More specifically, target devices need to support multi-touch, need to have at least accelerometers for six-DOF interactions, and WiFi for communication. Other factors such as screen size or weight influence usability, but are not technically limiting.

4 Implementation

Unlike most other approaches, our proposed interaction method is neither specifically tailored towards, nor implemented in, a single application. It is designed as a toolset for application or interaction designers wanting to include smart phone or tablet-based interactions in their applications. It can be included into a program as a library or plug-in. In our test implementation it has been added to the Vrui toolkit[3] [15] as a third party plug-in for and it can easily be included in any Vrui-based VR application.

4.1 Mobile Device

The prototype device-side application was developed in Objective-C for an Apple iPhone 4S and an iPad 2 (See [12] for potential problems when porting this to other platforms).

The device-side application implements the client component of the distributed architecture. On startup, it connects to an application-side server, found either via Service Discovery, or a manually entered host name. Upon connection, the client receives the application's menu structure and builds the menu's visual representation. If the application side requests orientation measurements, the client sends streaming orientation data at the maximal rate of 30 Hz to minimze lag. User interface events such as selection or quantification are reported asynchronously as they happen.

4.2 Extension Possibilities

The communication protocol is extensible and can be tailored toward a specific application should the need arise. Extensions are simple to implement on both sides and can make use of the already implemented features.

5 Conclusions

We have presented our eyes-free interaction method using consumer level smart phones and tablets. This method can be used effectively with applications for knowledge discovery and employs multi-touch and the phones sensors to allow full 3D interaction. Being eyes-free users can interact without interrupting their workflow to look at the input device. With its simple consistent design it can replace traditional menus and even whole dialog boxes. Our method is application-independent and can be used in other environments (e.g., Desktop) as well. Applications do not have to be altered to make use of this approach, but the prototype can be customized to include application-specific behavior.

A possible extension is device-based authentication. Devices can send public keys to the server on connection. Instead of a passphrase, the menu can be used

[3] `http://idav.ucdavis.edu/~okreylos/ResDev/Vrui`

to enter a combination of strokes serving as a password (similar to YAGP [8]). To prevent password sniffing, the communication can be encrypted using Transport Layer Security (TLS). The whole concept is also transferable to other application domains. When a multi-touch trackpad is available (as with most Macintosh computers) the menu can be controlled the same way. Unfortunately, the use of sensors to control rotation is then no longer possible.

The paradigm of our design is usable for a number of other application areas as well, such as desktop computers, tv sets, etc.

We plan do devise and perform a user study for our system, considering multiple applications and users from various disciplinary backgrounds (i.e., users applying out system to their domain-specific problems).

References

1. Accot, J., Zhai, S.: Beyond Fitts' law: models for trajectory-based HCI tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 295–302. ACM (1997)
2. Bailly, G., Lecolinet, E.: Flower menus: a new type of marking menu with large menu breadth, within groups and efficient expert mode memorization. In: AVI 2008 Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 15–22 (2008)
3. Ballagas, R., Rohs, M., Sheridan, J.G.: Sweep and Point and Shoot: Phonecam-based Interactions for Large Public Displays. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1200–1203. ACM (2005)
4. Baudisch, P., Good, N., Stewart, P.: Focus plus context screens: combining display technology with visualization techniques. In: Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology, UIST 2001, pp. 31–40. ACM, New York (2001)
5. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology 47(6), 381–391 (1954)
6. Francone, J., Bailly, G., Lecolinet, E., Mandran, N., Nigay, L.: Wavelet menus on handheld devices: stacking metaphor for novice mode and eyes-free selection for expert mode. In: Proceedings of the International Conference on Advanced Visual Interfaces, AVI 2010, pp. 173–180. ACM, New York (2010)
7. Francone, J., Bailly, G., Nigay, L., Lecolinet, E.: Wavelet Menus: A Stacking Metaphor for Adapting Marking Menus to Mobile Devices. In: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 2–5 (2009)
8. Gao, H., Guo, X., Chen, X., Wang, L., Liu, X.: YAGP: Yet Another Graphical Password Strategy. Computer Security Applications Conference, Annual, 121–129 (2008)
9. Guimbretiere, F., Winograd, T.: FlowMenu: Combining Command, Text, and Data Entry. In: Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology, pp. 213–216. ACM (2000)
10. Heidrich, F., Golod, I., Russell, P., Ziefle, M.: Device-free interaction in smart domestic environments. In: Proceedings of the 4th Augmented Human International Conference, AH 2013, pp. 65–68. ACM, New York (2013)

11. Holzinger, A., Searle, G., Peischl, B., Debevc, M.: An answer to who needs a stylus? on handwriting recognition on mobile devices. In: Obaidat, M.S., Sevillano, J.L., Filipe, J. (eds.) ICETE 2011. CCIS, vol. 314, pp. 156–167. Springer, Heidelberg (2012)
12. Holzinger, A., Treitler, P., Slany, W.: Making apps useable on multiple different mobile platforms: On interoperability for business application development on smartphones. In: Quirchmayr, G., Basl, J., You, I., Xu, L., Weippl, E. (eds.) CD-ARES 2012. LNCS, vol. 7465, pp. 176–189. Springer, Heidelberg (2012)
13. Jeon, S., Hwang, J., Kim, G.J., Billinghurst, M.: Interaction techniques in large display environments using hand-held devices. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology, VRST 2006, pp. 100–103. ACM, New York (2006)
14. Kin, K., Hartmann, B.: Two-handed marking menus for multitouch devices. ACM Transactions on Computer-Human Interaction (TOCHI) TOCHI Homepage Archive 18(2), 16:1–16:23 (2011)
15. Kreylos, O.: Environment-Independent VR Development. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Remagnino, P., Porikli, F., Peters, J., Klosowski, J., Arns, L., Chun, Y.K., Rhyne, T.-M., Monroe, L. (eds.) ISVC 2008, Part I. LNCS, vol. 5358, pp. 901–912. Springer, Heidelberg (2008)
16. Kurtenbach, G., Buxton, W.: The limits of expert performance using hierarchic marking menus. In: Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems, CHI 1993, pp. 482–487. ACM, New York (1993)
17. Kurtenbach, G., Buxton, W.: User learning and performance with marking menus. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence, CHI 1994, pp. 258–264. ACM, New York (1994)
18. Lepinski, G., Grossman, T.: The design and evaluation of multitouch marking menus. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, pp. 2233–2242 (2010)
19. Madhavapeddy, A., Scott, D., Sharp, R.: Using Camera-Phones to Enhance Human-Computer Interaction. In: Proceedings of Ubiquitous (2004)
20. Mankoff, J., Abowd, G.D.: Cirrin: a word-level unistroke keyboard for pen input. In: Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology, UIST 1998, pp. 213–214. ACM, New York (1998)
21. Perlin, K.: Quikwriting: continuous stylus-based text entry. In: Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology, UIST 1998, pp. 215–216. ACM, New York (1998)
22. Pook, S., Lecolinet, E., Vaysseix, G., Ura, E.C., Bp, M.: Control Menus: Execution and Control in a Single Interactor. In: CHI 2000 Extended Abstracts on Human Factors in Computing Systems, pp. 263–264 (April 2000)
23. Thelen, S., Meyer, J., Ebert, A., Hagen, H.: A 3D Human Brain Atlas. Modelling the Physiological Human, 173–186 (2009)
24. Venolia, D., Neiberg, F.: T-Cube: a fast, self-disclosing pen-based alphabet. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence, CHI 1994, pp. 265–270. ACM, New York (1994)

On Visual Analytics and Evaluation in Cell Physiology: A Case Study

Fleur Jeanquartier and Andreas Holzinger

Research Unit Human-Computer Interaction, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz
{f.jeanquartier,a.holzinger}@hci4all.at

Abstract. In this paper we present a case study on a visual analytics (VA) process on the example of cell physiology. Following the model of Keim, we illustrate the steps required within an exploration and sense-making process. Moreover, we demonstrate the applicability of this model and show several shortcomings in the analysis tools functionality and usability. The case study highlights the need for conducting evaluation and improvements in VA in the domain of biomedical science. The main issue is the absence of a complete toolset that supports all analysis tasks including the many steps of data preprocessing as well as end-user development. Another important issue is to enable collaboration by creating the possibility of evaluating and validating datasets, comparing it with data of other similar research groups.

Keywords: visual analytics, evaluation of visualization, human computer interaction, biomedical science.

1 Introduction

From the first data analysis attempts to exploratory data analysis, up to information visualization, today we are facing the possibilities of visual analytics (VA). With VA several analysis processes may be transformed and become more effective and efficient through integrating automated analysis results and reasoning [1]. There is ongoing research in a variety of application areas ranging from document analysis over network security to molecular biology. Applying visual analysis techniques within these areas bring up certain limitations [2]. Dealing with the complexity of biological data requires sophisticated visualization technologies. Prominent examples of visualization for exploration and analysis in the domain of biology come from systems biology and include, among many others, the visualization of biological networks and omics data [3] such as protein structures [4], visual analysis of gene expression data [5], but also visual analysis of cell signaling networks [6]. For populating such databases for network analysis biologists also deal with basic research in cell physiology.

In Fig. 1 we see a slightly modified version of the VA Process, first described by [7]. According to Keim, humans have to be included early in the data analysis

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 495–502, 2013.

process. By using their background knowledge and being supported by processing, transformation and visualization tools the analysis process eventually brings up new insight. We illustrate the mapping of a VA process during one example cell physiological experiment. Life scientists especially in the domain of biomedical science may struggle with the fact, that the process starts with a first data analysis. As for the observed work process later described in the Section 2 the domain expert also started with describing the hypothesis and then choosing suitable materials and methods for data acquisition. According to [7] input for the data sets used in the VA process are of heterogeneous nature and can be results from scientific experiments. Therefore we included the prior results as input for the feedback loop. The hypothesis may be formed by a preceding exploration. The domain expert makes use of knowledge gained by preceding work.

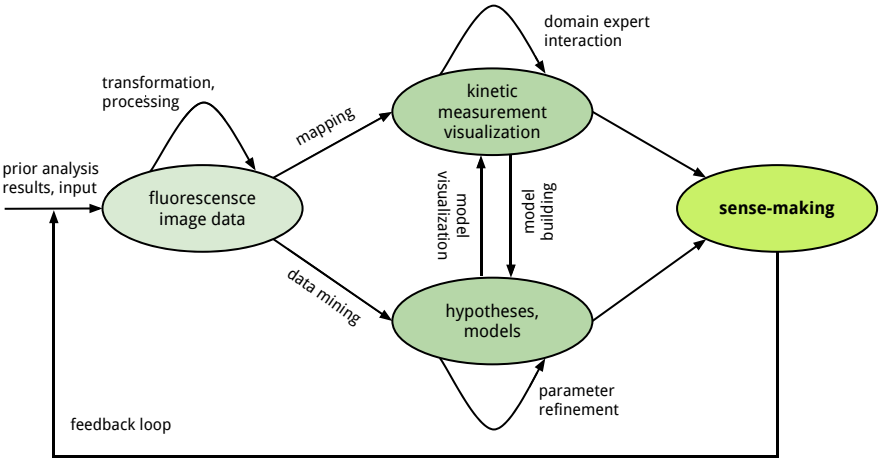


Fig. 1. Adapted Version of Keim's VA Process for the Application In Cell Physiology

Human computer interaction (HCI) and knowledge discovery (KDD) along with biomedical informatics are of increasing importance to effectively make sense out of data [8]. Biologists can benefit from a data deluge with the means of an integrated visualization approach, however, conducting evaluation and improving toolsets are still required to overcome certain hurdles on the way to new insights [9]. A domain expert as analyst often works alone while analysing data sets facing many problems, some of that have been illustrated by [10]. To foster sense-making and insights in VA systems it is essential to conduct studies and determine how people are using such systems [11, 12]. Case studies as field studies are a common approach to evaluating VA systems [13]. Qualitative evaluation such as observational studies can be conducted in a more realistic setting and allow improved understanding of existing practices for analysis and environmental constraints [14].

Consequently, we describe an observational study of a domain expert in cell physiology to present the current practice of VA in this domain.

2 Observation

The user, a domain expert within biomedical science, is part of the visual analysis and KDD process of a group of researchers dealing with cell physiology experiments. We accompanied the domain expert while investigating and analysing a set of experiments' results and observed the expert's analysis work. The analysis process includes visual analysis as part of the data processing, data analysis and KDD process as well as visual communication for dissemination.

A fluorescent biosensor [15] measures the concentration of certain molecules within cellular compartments. Fluorescent biosensors can be used for monitoring various processes and analytes such as metabolites, ions, target localization, gene expression and physiological relevant changes within subcellular regions [16]. The biosensor allows to quantify variations in concentration or localization of the specific analyte within the cell by a change in fluorescence intensity. This quantification is further visualized as intensity signal over time in terms of kinetic curves. By that method, data in hundreds of columns and rows is recorded and has to be processed further. In summary, this method provides the measurement of biological signaling dynamics in vivo.

Experiments start with monitoring kinetics in signal transduction. The signal represents the fluorescence intensity [17]. First of all sequences of high-resolution fluorescent imaging of cells are acquired to capture dynamic changes. This action takes place in the lab's dark room. Fluorescence images are captured by a digital camera incorporating a CCD detector, connected to the fluorescence microscope. A commercial bioimaging software is used to communicate with the hardware, translating recorded signals to raw data. The software also provides some data/image processing functionality. Once the measurements are complete, the analysis process continues with data processing and image analysis. Noise (such as background lights within the dark room) reduction of images is supported by a ratio function. The domain expert marks specific regions of interest within the cell in order to monitor biological activites in healthy and pathological cells. Image segmentation is done manually insofar as the domain expert manually selects specific regions of interest on the image data for further comparison and analysis. Hence, regions of interest as polygon shapes are placed on every raw source to display the intensity value. The evaluation of whether the data and to what extent is accurate is done by manually comparing specific regions with a background region. The software allows the scientist to explore the data only in a very limited way. For not occupying the lab's dark room workplace for the time-consuming tasks of data processing and analysis, the expert moves to another workplace outside the dark room. Consequently, when the domain expert believes, that the data is sufficient, the raw data is exported to a commercial spreadsheet computation software via CSV for further processing and analysis.

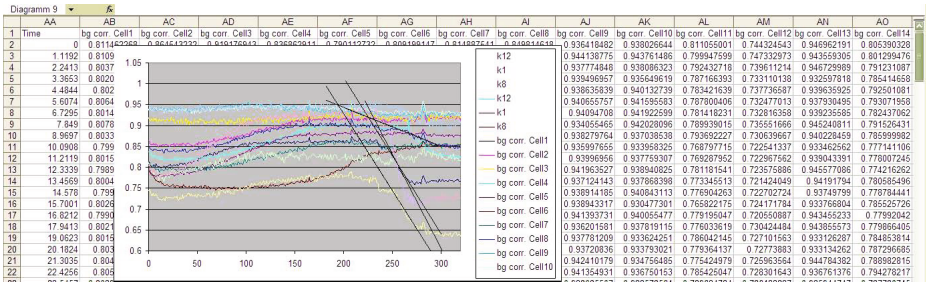


Fig. 2. First visual analysis of intensity signal over time

The domain expert creates a first visualization (compare Fig. 2) of the data, describing the kinetic changes in specific groups of healthy and pathological cells. This task is done semi-automatically by end-user development [18]. The following visual analysis shows, that the data has to be further filtered, corrected and transformed and finally improved in terms of readability and to be visualizable for the task of dissemination. It is up to the domain expert and the implicit knowledge of models, developed by the group of researchers within the lab, which transformation and manipulations are considered to be appropriate. Some of the processing tasks are automatic and some are again manual. The domain expert uses several tools for the various tasks and switches between them while advancing in the analysis process. While the process itself is occasionally being discussed in group, several smaller but complex actions are double checked by colleagues. Both the experiments and the visual analysis process are repeated many times until certain "surprising" [19] results get visible. This repetitive approach to gain new insights, also known as explorative data analysis (EDA), supports the process illustrated in Fig. 1 as it consists of a feedback loop. Finally, when the visual analysis results show surprising effects, the domain experts concludes with the dissemination (see Fig. 3) of the results, again with the means of visualization. The final visualizations are again being iteratively improved.

By further discussing the case study's process and comparing it with the VA process, we try to outline certain issues when dealing with the evaluation of scientific visualizations.

3 Discussion

The case study shows, that there are analysis processes in biomedical science which embody VA as a lived approach. At the same time, the case study also shows the need for improvements regarding HCI and end-user development. Experts in this domain are using their domain knowledge in combination with both automatic and visual analysis together, but need to be guided by computer science experts to improve the choice of tools that are used. There are certain tasks still done manually that could be automated or at least semi-automated, using

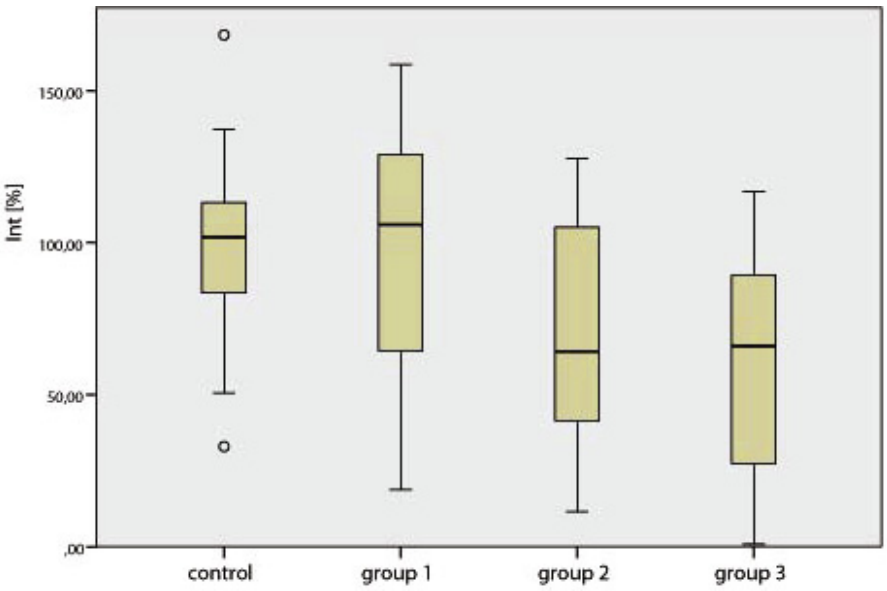


Fig. 3. Visualization of the kinetic parameters measured within the groups of healthy control to pathological cells for dissemination

the right tool, such as image segmentation, noise reduction as well as post-processing up to creating a set of fitting visualizations for dissemination. For instance, there already exist attempts to automatically estimate suitable background in fluorescence imaging [20] and automatic localization of cell nuclei [21]. The case study supports the statement, proposed by VA, that automated analysis often speeds up analysis tasks. It also shows that communication through visual representations is used for the dissemination of research results (see Fig. 3). Therefore, the case study shows, that visual analysis plays an important role for gaining new insights and dissemination. However, the case study also highlights that certain evaluation tasks are missing.

Lessons Learned include that there is a gap in free exploration of data and information due to the lacking usability and interaction possibilities of the tools used. The researchers within this field of study state, that they do not know about powerful VA solutions. At the same time, they are facing certain restrictions that hinder them to cooperate fully with computer science experts. During the observation the domain expert made complaints about shortcomings in the analysis tools' functionality and usability. Many tasks have to be repeated, not only due to data inconsistency, but also because most tools are hardly fault-tolerant and lack in supporting the user in certain data preprocessing steps as well as in the post-processing such as choosing the right visualization technique and improving the visualization's readability.

There are several possibilities to improve the end-user development and to minimize the interaction junk [22] within the observed process, such as simplifying

the creation of the effect curves for both visual analysis as well as dissemination. Furthermore, the visualizations in use are still limited to curve diagrams and bar charts. Alternative visualization metaphors such as multi-variate data visualizations [23] allow scientists to explore the data and it's various dimensions in other ways and may highlight certain effects that are not visible within the current effect curves.

The discussion after the observation further included improvements and suggestions to support the whole VA process. Due to the reason that both data as well as study results are confidential we are not allowed to go into detail in this respect. However, we already communicate general aspects of HCI and KDD and present general suggestions for improving VA within this domain. The domain expert agreed that there are several possibilities how evaluation could be integrated to support VA. Lam et al. already list some fitting evaluation goals and questions within the VDAR- and the CTV scenario [13]. However, the very idea of discussing the visual analysis process with a domain expert in HCI already brought up certain shortcomings within the visual analysis work. Suggestions include: Evaluating the dataset, comparing it to datasets of other similar groups of researchers, would help validating specific models as well as techniques and speed up the analysis work. Moreover, enabling and facilitating collaboration supports scientific problem solving [24]. The researchers in the group also agree on the fact, that evaluating software in use and furthermore, having the possibility to improve and extend the tools functionality would improve their daily research tasks. Incooperating the many steps of data examination and preprocessing into a single tool would be highly appreciated. The case study highlights the need for conducting evaluation and improvements in VA in the domain of biomedical science.

4 Conclusion

Every day scientists in many sub domains of life sciences such as biomedical science are facing the challenging task of VA with the goal of reaching new insights. Life scientists may benefit from a data deluge with the means of an integrated visualization approach. However, conducting evaluation and improving certain toolsets for exploratory data analysis and end-user development are prominent challenges on the way to new insights.

We described an observational study of VA in cell physiology. We compared the process to Keim's VA process. The case study shows, that there are analysis processes in biomedical science which embody VA. Further studies may include additional practice of VA related analysis work of various other approaches in biomedical science. The observation highlights the need for conducting evaluation and improvements in VA in the domain of biomedical science. We suggested evaluation possibilities and further noted challenges regarding its' application for visualization in life sciences. Among others, suggestions include incooperation and improvement of support for developing visualization in regard to analysis.

References

[1] Keim, D.A., Kohlhammer, J., Ellis, G., Mansmann, F.: Mastering The Information Age-Solving Problems with Visual Analytics. Florian Mansmann (2010)

[2] Kohlhammer, J., Keim, D., Pohl, M., Santucci, G., Andrienko, G.: Solving Problems with Visual Analytics. Procedia Computer Science 7, 117–120 (2011)

[3] Gehlenborg, N., O'Donoghue, S.I., Baliga, N.S., Goesmann, A., Hibbs, M.A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., et al.: Visualization of omics data for systems biology. Nature Methods 7, S56–S68 (2010)

[4] Doncheva, N.T., Assenov, Y., Domingues, F.S., Albrecht, M.: Topological analysis and interactive visualization of biological networks and protein structures. Nature Protocols 7(4), 670–685 (2012)

[5] Lex, A., Streit, M., Kruijff, E., Schmalstieg, D.: Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context. In: 2010 IEEE Pacific Visualization Symposium (PacificVis), pp. 57–64. IEEE (2010)

[6] Berger, S.I., Iyengar, R., Maayan, A.: Avis: Ajax viewer of interactive signaling networks. Bioinformatics 23(20), 2803–2805 (2007)

[7] Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual analytics: Scope and challenges. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 76–90. Springer, Heidelberg (2008)

[8] Holzinger, A.: On Knowledge Discovery and interactive intelligent visualization of biomedical data: Challenges in Human–Computer Interaction & Biomedical Informatics. In: DATA-International Conference on Data Technologies and Applications, pp. 5–16 (2012)

[9] Donoghue, S.I.O., Gavin, A.-c., Gehlenborg, N., Goodsell, D.S., Hériché, J.-k., Nielsen, C.B., North, C., Olson, A.J., Procter, J.B., Shattuck, D.W., Walter, T., Wong, B.: Visualizing biological data - now and in the future. Nature Publishing Group 7(3), S2–S4 (2010)

[10] Wong, B.L.W., Xu, K., Holzinger, A.: Interactive Visualization for Information Analysis in Medical Diagnosis. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 109–120. Springer, Heidelberg (2011)

[11] Kang, Y.A., Görg, C., Stasko, J.: How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation. IEEE Transactions on Visualization and Computer Graphics 17(5), 570–583 (2010)

[12] Wong, P.C., Shen, H.-W., Johnson, C.R., Chen, C., Ross, R.B.: The Top 10 Challenges in Extreme-Scale Visual Analytics. IEEE Computer Graphics and Applications 32(4), 63–67 (2012)

[13] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical Studies in Information Visualization: Seven Scenarios.. IEEE Transactions on Visualization and Computer Graphics 18(9), 1–18 (2011)

[14] Carpendale, S.: Evaluating information visualizations. In: Kerren, A., Stasko, J.T., Fekete, J.-D., North, C. (eds.) Information Visualization. LNCS, vol. 4950, pp. 19–45. Springer, Heidelberg (2008)

[15] Morris, M.C.: Fluorescent biosensors of intracellular targets from genetically encoded reporters to modular polypeptide probes. Cell Biochemistry and Biophysics 56(1), 19–37 (2010)

[16] Okumoto, S., Jones, A., Frommer, W.B.: Quantitative imaging with fluorescent biosensors. Annual Review of Plant Biology 63, 663–706 (2012)

[17] Mehta, S., Zhang, J.: Reporting from the field: genetically encoded fluorescent reporters uncover signaling dynamics in living biological systems.. Annual Review of Biochemistry 80, 375–401 (2011)

[18] Lieberman, H., Paternò, F., Wulf, V.: End user development, vol. 9. Springer (2006)
[19] Beale, R.: Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing. International Journal of Human-Computer Studies 65(5), 421–433 (2007)
[20] Chen, T.-W., Lin, B.-J., Brunner, E., Schild, D.: In situ background estimation in quantitative fluorescence imaging.. Biophysical Journal 90(7), 2534–2547 (2006)
[21] Song, Y., Cai, W., Huang, H., Wang, Y., Feng, D.D., Chen, M.: Region-based progressive localization of cell nuclei in microscopic images with data adaptive modeling. BMC Bioinformatics 14(1), 173 (2013)
[22] Endert, A., North, C.: Interaction junk. In: Proceedings of the 2012 BELIV Workshop on Beyond Time and Errors - Novel Evaluation Methods for Visualization - BELIV 2012, pp. 1–3. ACM Press, New York (2012)
[23] Fuchs, R., Hauser, H.: Visualization of Multi-Variate Scientific Data. Computer Graphics Forum 28(6), 1670–1690 (2009)
[24] Good, B.M., Su, A.I.: Games with a scientific purpose. Genome Biology 12(12), 135 (2011)

Author Index