Ana Fred · Jan Dietz
David Aveiro · Kecheng Liu
Jorge Bernardino · Joaquim Filipe (Eds.)

Communications in Computer and Information Science 914

Knowledge Discovery, Knowledge Engineering and Knowledge Management

8th International Joint Conference, IC3K 2016
Porto, Portugal, November 9–11, 2016
Revised Selected Papers

Springer

Communications in Computer and Information Science 914

More information about this series at http://www.springer.com/series/7899

Ana Fred · Jan Dietz
David Aveiro · Kecheng Liu
Jorge Bernardino · Joaquim Filipe (Eds.)

Knowledge Discovery, Knowledge Engineering and Knowledge Management

8th International Joint Conference, IC3K 2016
Porto, Portugal, November 9–11, 2016
Revised Selected Papers

Springer

Editors

Ana Fred
Instituto de Telecomunicações
Lisbon, Portugal

Jan Dietz
Department of Software Technology
Delft University of Technology
Voorburg, Zuid-Holland, The Netherlands

David Aveiro
Faculty of Exact Sciences and Engineering
University of Madeira
Funchal, Portugal

and

Madeira Interactive Technologies Institute
Funchal, Portugal

Kecheng Liu
Henley Business School
University of Reading
Reading, UK

Jorge Bernardino
University of Coimbra
Coimbra, Portugal

Joaquim Filipe
Instituto Politecnico de Setúbal (IPS)
Setúbal, Portugal

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-99700-1 ISBN 978-3-319-99701-8 (eBook)
https://doi.org/10.1007/978-3-319-99701-8

Library of Congress Control Number: 2018953365

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The present book includes extended and revised versions of a set of selected papers from the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), held in Porto, Portugal, during November 9–11, 2016.

IC3K 2016 received 186 paper submissions from 46 countries, of which 9% were included in this book. The papers were selected by the event chairs and their selection was based on a number of criteria that include the classifications and comments provided by the Program Committee members, the session chairs' assessment, and also the program chairs' global view of all papers included in the technical program. The authors of selected papers were then invited to submit a revised and extended version of their papers with an addition of at least 30% novel material.

The purpose of IC3K is to bring together researchers, engineers, and practitioners on the areas of knowledge discovery, knowledge engineering, and knowledge management. IC3K is composed of three co-located conferences (KDIR, KEOD, and KMIS), each specialized in at least one of the aforementioned main knowledge areas.

The papers selected to be included in this book contribute to the understanding of relevant topics. We selected three papers focusing on machine learning, two of them with applications to data mining and one related to natural language processing; in an adjacent area we selected four other papers addressing text mining or data mining in detail. In the area of knowledge engineering, there are three papers addressing ontology engineering and knowledge representation and one focusing on conceptual software design. Finally, four papers addressing intelligent information systems, in particular communication, collaboration, and information sharing, and three papers focusing on knowledge management and enterprise systems complete the collection for this book of a total of 18 papers.

We would like to thank all the authors for their contributions and also the reviewers who helped ensure the quality of this publication.

November 2016

Ana Fred
Jan Dietz
David Aveiro
Kecheng Liu
Jorge Bernardino
Joaquim Filipe

The original version of the book was revised: For detailed information please see Correction. The correction to the book is available at https://doi.org/10.1007/978-3-319-99701-8_19

Organization

Conference Chair

Joaquim Filipe	Polytechnic Institute of Setúbal/INSTICC, Portugal

Program Co-chairs

KDIR

Ana Fred	Instituto de Telecomunicações, Portugal

KEOD

Jan Dietz	Delft University of Technology, The Netherlands
David Aveiro	University of Madeira/Madeira Interactive Technologies Institute, Portugal

KMIS

Kecheng Liu	University of Reading, UK
Jorge Bernardino	Polytechnic Institute of Coimbra, ISEC, Portugal

KDIR Program Committee

Ayan Acharya	CognitiveScale Inc., USA
Dirk Ahlers	NTNU, Norway
Amir Ahmad	United Arab Emirates University, UAE
Mayer Aladjem	Ben-Gurion University of the Negev, Israel
Tiago Almeida	Federal University of São Carlos, Brazil
Eva Armengol	IIIA CSIC, Spain
Zeyar Aung	Masdar Institute of Science and Technology, UAE
Vladan Babovic	National University of Singapore, Singapore
Niranjan Balasubramanian	University of Massachusetts Amherst, USA
Vladimir Bartik	Brno University of Technology, Czech Republic
Márcio Basgalupp	Universidade Federal de São Paulo, Brazil
Gloria Bordogna	CNR, National Research Council, Italy
Luis M. de Campos	University of Granada, Spain
Huiping Cao	New Mexico State University, USA
Jesús Ariel Carrasco-Ochoa	INAOE, Mexico
Maria Catalan	Jaume I University, Spain
Michelangelo Ceci	University of Bari, Italy
Sunandan Chakraborty	New York University, USA
Chien-Chung Chan	University of Akron, USA
Keith C. C. Chan	Hong Kong Polytechnic University, SAR China

Praveen Chandar	IBM Research, USA
Lijun Chang	UNSW, Australia
Chien Chin Chen	National Taiwan University College of Management, Taiwan
Zhiyuan Chen	University of Maryland Baltimore County, USA
Kun-Ta Chuang	National Cheng Kung University Taiwan, Taiwan
Patrick Marques Ciarelli	Universidade Federal do Espírito Santo, Brazil
Carlo Combi	Università degli Studi di Verona, Italy
Paulo Cortez	University of Minho, Portugal
Ingemar Cox	University of Copenhagen, Denmark
Bruce Croft	University of Massachusetts Amherst, USA
Thanh-Nghi Do	Can Tho University, Vietnam
Antoine Doucet	University of La Rochelle, France
Markus Endres	University of Augsburg, Germany
Floriana Esposito	Università degli Studi di Bari Aldo Moro, Italy
Iaakov Exman	The Jerusalem College of Engineering, Azrieli, Israel
Dan Feldman	MIT, Israel
Elisabetta Fersini	University of Milano-Bicocca, Italy
Philippe Fournier-Viger	University of Moncton, Canada
Fabrício Olivetti de França	Universidade Federal do ABC, Brazil
Ana Fred	Instituto de Telecomunicações, Portugal
Panorea Gaitanou	Ionian University, Greece
William Gasarch	University of Maryland, USA
Michael Gashler	University of Arkansas, USA
Susan Gauch	University of Arkansas, USA
Judith Gelernter	National Institute of Standards and Technology, USA
Rosario Girardi	UFMA, Brazil
Rosalba Giugno	University of Catania, Italy
Lorraine Goeuriot	Joseph Fourier University, France
Francesco Gullo	UniCredit R&D, Italy
Jennifer Harding	Loughborough University, UK
Samuel Huston	Google Inc., USA
Beatriz de la Iglesia	University of East Anglia, UK
Jiepu Jiang	University of Massachusetts Amherst, USA
Mouna Kamel	IRIT, France
Mehmed Kantardzic	University of Louisville, USA
Ron Kenett	Samuel Neaman Institute, Technion, Israel
Yun Sing Koh	University of Auckland, New Zealand
Margita Kon-Popovska	Ss Cyril and Methodius University, Macedonia, The Former Yugoslav Republic of
Ioannis Konstantinou	NTUA, Greece
Steven Kraines	University of Tokyo, Japan
Ralf Krestel	Hasso-Plattner-Institute, Germany
Vasileios Lampos	University College London, UK
Hagen Langer	University of Bremen, Germany
Nuno Lau	Universidade de Aveiro, Portugal
Anne Laurent	LIRMM, Montpellier University, France

Seamus Lawless	Trinity College Dublin, Ireland
Carson K. Leung	University of Manitoba, Canada
Johannes Leveling	Elsevier, The Netherlands
Chun Hung Li	Hong Kong Baptist University, SAR China
Chun-Wei Lin	Western Norway University of Applied Sciences, Norway
Xia Lin	Drexel University, USA
Michel Liquiere	University of Montpellier II, France
Berenike Litz	Attensity Corporation, USA
Giovanni Livraga	Università degli Studi di Milano, Italy
Rafael Berlanga Llavori	Jaume I University, Spain
Miguel Angel Guevara Lopez	University of Minho, Portugal
Alicia Troncoso Lora	Pablo de Olavide University of Seville, Spain
Devignes Marie-Dominique	LORIA, CNRS, France
J. Francisco Martínez-Trinidad	Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico
Pietro Michiardi	EURECOM, France
Bamshad Mobasher	DePaul University, USA
Misael Mongiovi	Università di Catania, Italy
Stefania Montani	Piemonte Orientale University, Italy
Eduardo F. Morales	INAOE, Mexico
Yashar Moshfeghi	University of Glasgow, UK
Josiane Mothe	Université de Toulouse, France
Xia Ning	IUPUI, USA
Mitsunori Ogihara	University of Miami, USA
Neil O'Hare	Yahoo, USA
Elias Oliveira	Universidade Federal do Espirito Santo, Brazil
José Luis Oliveira	Universidade de Aveiro, Portugal
Márcia Oliveira	Universidade Federal do Espírito Santo, Brazil
Colm O'Riordan	NUI, Galway, Ireland
Rifat Ozcan	Turgut Ozal University, Turkey
Sarala Padi	National Institute of Standards and Technology, USA
Rui Pedro Paiva	University of Coimbra, Portugal
Krzysztof Pancerz	University of Rzeszow, Poland
Gaurav Pandey	Icahn School of Medicine at Mount Sinai, USA
NhatHai Phan	University of Oregon, USA
Karen Pinel-Sauvagnat	IRIT, France
Alberto Adrego Pinto	University of Porto, Portugal
Gianvito Pio	Università degli Studi di Bari Aldo Moro, Italy
Luigi Pontieri	National Research Council, Italy
Alan L. Porter	Georgia Institute of Technology, USA
Ronaldo Prati	Universidade Federal do ABC, Brazil
Marcos Gonçalves Quiles	Federal University of Sao Paulo, Brazil
Chiara Renso	ISTI-CNR, Italy
Carolina Ruiz	WPI, USA

Henryk Rybinski	Warsaw University of Technology, Poland
Ovidio Salvetti	National Research Council, Italy
Vasilis Samoladas	Technical University of Crete, Greece
Ralf Schenkel	University of Trier, Germany
Filippo Sciarrone	Roma Tre University, Italy
Zhongzhi Shi	Chinese Academy of Sciences, China
Andrzej Sluzek	Khalifa University, UAE
Minseok Song	Pohang University of Science and Technology, South Korea
Kostas Stefanidis	ICS-FORTH, Greece
Ulrich Thiel	Fraunhofer Gesellschaft, Germany
James Thom	RMIT University, Australia
I-Hsien Ting	National University of Kaohsiung, Taiwan
Kar Ann Toh	Yonsei University, South Korea
Predrag Tosic	Washington State University, USA
Domenico Ursino	Università Politecnica delle Marche, Italy
Nina Wacholder	Rutgers University, USA
Bingsheng Wang	Google, USA
Jiabing Wang	South China University of Technology, China
Peiling Wang	University of Tennessee Knoxville, USA
Leandro Krug Wives	Universidade Federal do Rio Grande do Sul, Brazil
Yanghua Xiao	Fudan University, China
Ce Zhang	Stanford University, USA
Yi Zhang	University of Technology Sydney, Australia
Wei Zhou	ESCP Europe, France

KDIR Additional Reviewers

Pengwei Hu	Hong Kong Polytechnic University, SAR China
Sarala Padi	NIST, USA
Yuhao Yang	ZL Technologies, USA

KEOD Program Committee

Raian Ali	Bournemouth University, UK
Carlo Allocca	FORTH Research Institute, University of Crete, Greece
Frederic Andres	Research Organization of Information and Systems, Japan
Francisco Antunes	Institute of Computer and Systems Engineering of Coimbra and Beira Interior University, Portugal
David Aveiro	University of Madeira/Madeira Interactive Technologies Institute, Portugal
Claudio de Souza Baptista	Universidade Federal de Campina Grande, Brazil
Teresa M. A. Basile	Università degli Studi di Bari, Italy
Sonia Bergamaschi	University of Modena and Reggio Emilia, Italy
Gerhard Budin	University of Vienna, Austria
Radek Burget	Brno University of Technology, Czech Republic

Doina Caragea	Kansas State University, USA
Tsan-Ming Choi	Hong Kong Polytechnic University, SAR China
Davide Ciucci	University of Milano-Bicocca, Italy
João Paulo Costa	Institute of Computer and Systems Engineering of Coimbra, Portugal
Christophe Cruz	CNRS, France
Erdogan Dogdu	TOBB University of Economics and Technology, Turkey
Pierpaolo D'Urso	Università di Roma La Sapienza, Italy
John Edwards	Aston University, UK
Henrik Eriksson	Linköping University, Sweden
Anna Fensel	STI Innsbruck, University of Innsbruck, Austria
Dieter A. Fensel	University of Innsbruck, Austria
Jesualdo Tomás Fernández-Breis	University of Murcia, Spain
George Giannakopoulos	NCSR Demokritos, Greece
Yoan Gutiérrez	University of Alicante, Spain
Mamoun Abu Helou	Al-Istiqlal University, Palestinian Territory, Occupied
Christopher Hogger	Imperial College London, UK
Achilles Kameas	Hellenic Open University, Greece
Dimitris Kanellopoulos	University of Patras, Greece
Sarantos Kapidakis	Ionian University, Greece
Pinar Karagoz	METU, Turkey
Kouji Kozaki	Osaka University, Japan
Antoni Ligeza	AGH University of Science and Technology, Poland
Elena Lloret	University of Alicante, Spain
Paulo Maio	Polytechnic of Porto, Portugal
Luca Mazzola	Lucerne University of Applied Sciences, Switzerland
Rocio Abascal Mena	Universidad Autónoma Metropolitana, Cuajimalpa, Mexico
John-Jules Meyer	Utrecht University, The Netherlands
Riichiro Mizoguchi	Japan Advanced Institute of Science and Technology, Japan
Andres Montoyo	University of Alicante, Spain
Claude Moulin	JRU CNRS Heudiasyc, University of Compiègne, France
Ana Maria Moura	National Laboratory of Scientific Computing, Brazil
Phivos Mylonas	Ionian University, Greece
William Nelson	Devry University, USA
Erich Neuhold	University of Vienna, Austria
Jørgen Fischer Nilsson	Technical University of Denmark, Denmark
Femke Ongenae	Ghent University - imec, Belgium
Matteo Palmonari	University of Milano-Bicocca, Italy
Jiajie Peng	Northwestern Polytechnical University, China
Carlos Periñán-Pascual	Universidad Politécnica de Valencia, Spain
Dimitris Plexousakis	FORTH, Greece

Mihail Popescu	University of Missouri-Columbia, USA
Nives Mikelic Preradovic	University of Zagreb, Croatia
Amar Ramdane-Cherif	University of Versailles St Quentin en Yvelines, France
Domenico Redavid	Consorzio Interuniversitario Nazionale per L’Informatica, Italy
Thomas Risse	University Library Johann Christian Senckenberg, Germany
Oscar Rodríguez Rocha	Inria Sophia Antipolis-Méditerranée, France
Colette Rolland	Université de Paris 1 Panthèon Sorbonne, France
Anisa Rula	University of Milano-Bicocca, Italy
Inès Saad	ESC Amiens, France
Fabio Sartori	University of Milano-Bicocca, Italy
Sagar Sen	Simula Research Laboratory, Norway
Antonio Lucas Soares	FEUP and INESC TEC, Portugal
Chuan Sun	NYC Data Science Academy, USA
Cesar Augusto Tacla	Federal University of Technology in Parana, Brazil
Gheorghe Tecuci	George Mason University, USA
Orazio Tomarchio	University of Catania, Italy
Shengru Tu	University of New Orleans, USA
Petr Tucnik	University of Hradec Kralove, Czech Republic
Yannis Tzitzikas	University of Crete, Greece
Iraklis Varlamis	Harokopio University of Athens, Greece
Bruno Volckaert	Ghent University - imec, Belgium
Toyohide Watanabe	Nagoya Industrial Science Research Institute, Japan
Yue Xu	Queensland University of Technology, Australia
Gian Piero Zarri	Sorbonne University, France
Jinglan Zhang	Queensland University of Technology, Australia
Ying Zhao	Naval Postgraduate School, USA
Qiang Zhu	University of Michigan, Dearborn, USA

KEOD Additional Reviewers

Sarra Bouzayane	University of Picardie Jules Verne Amiens France, France
Ntina Kakali	Panteion University Library and Information Centre, Greece
Blerina Spahiu	University of Milano-Bicocca, Italy
Stijn Verstichel	Ghent University - iMinds, Belgium

KMIS Program Committee

Marie-Helene Abel	University of Compiègne, France
Miriam C. Bergue Alves	Institute of Aeronautics and Space, Brazil
Rangachari Anand	IBM T. J. Watson Research Center, USA
Carlos Alberto Malcher Bastos	Universidade Federal Fluminense, Brazil

Jorge Bernardino	Polytechnic Institute of Coimbra, Portugal
Kelly Braghetto	University of São Paulo, Brazil
Roger Chiang	University of Cincinnati, USA
Dickson K. W. Chiu	Dickson Computer Systems, Hong Kong, SAR China
Byron Choi	Hong Kong Baptist University, SAR China
Ritesh Chugh	Central Queensland University, Australia
Marie-Christine Fauvet	Université Grenoble Alpes, France
Joan-Francesc Fondevila-Gascón	Centre d'Estudis sobre el Cable, UPF, URL, UdG (EU Mediterrani) and UOC, Spain
Yiwei Gong	Wuhan University, China
Anna Goy	University of Turin, Italy
Renata Guizzardi	Universidade Federal do Espírito Santo, Brazil
Anne Håkansson	KTH, Sweden
Jennifer Harding	Loughborough University, UK
María V. Hurtado	Universidad de Granada, Spain
Maria-Eugenia Iacob	University of Twente, The Netherlands
Anca Daniela Ionita	University Politehnica of Bucharest, Romania
Radoslaw Katarzyniak	Wroclaw University of Science and Technology, Poland
Mieczyslaw Klopotek	Polish Academy of Sciences, Poland
Veit Koeppen	Otto von Guericke University of Magdeburg, Germany
Helmut Krcmar	Technical University of Munich, Germany
Tri A. Kurniawan	Brawijaya University, Indonesia
Dominique Laurent	Cergy-Pontoise University, ENSEA, France
Michael Leyer	University of Rostock, Germany
Kecheng Liu	University of Reading, UK
Lin Liu	Tsinghua University, China
Heide Lukosch	Delft University of Technology, The Netherlands
Fabrizio Maria Maggi	University of Tartu, Estonia
Federica Mandreoli	University of Modena and Reggio Emilia, Italy
Nada Matta	University of Technology of Troyes, France
Rodney McAdam	University of Ulster, UK
Christine Michel	INSA-Lyon, Laboratoire LIRIS, France
Michele M. Missikoff	ISTC-CNR, Italy
Normen Müller	Safeplace at Protection One GmbH, Germany
Fabio Nonino	Università degli Studi di Roma La Sapienza, Italy
Andreas Oberweis	Karlsruhe Institute of Technology, Germany
Jonice Oliveira	UFRJ, Brazil
Silvia Dallavalle de Pádua	Universidade de São Paulo, Brazil
Wilma Penzo	University of Bologna, Italy
Erwin Pesch	University Siegen, Germany
Filipe Portela	University of Minho, Portugal
Arkalgud Ramaprasad	University of Illinois at Chicago, USA
Marina Ribaudo	Università di Genova, Italy
Colette Rolland	Université de Paris 1 Panthèon Sorbonne, France
Ana Roxin	University of Burgundy, France

Umar Ruhi	University of Ottawa, Canada
Masaki Samejima	Osaka University, Japan
Christian Seel	University of Applied Sciences Landshut, Germany
Mukhammad Andri Setiawan	Universitas Islam Indonesia, Indonesia
Tijs Slaats	University of Copenhagen, Denmark
Malgorzata Sterna	Poznan University of Technology, Poland
Jeff K.T. Tang	Caritas Institute of Higher Education, Hong Kong, SAR China
Shu-Mei Tseng	I-SHOU University, Taiwan
Jose Luis de la Vara	Carlos III University of Madrid, Spain
Rosina Weber	Drexel University, USA
Martin Wessner	Darmstadt University of Applied Sciences, Germany
Uffe K. Wiil	University of Southern Denmark, Denmark
Leandro Krug Wives	Universidade Federal do Rio Grande do Sul, Brazil
Rafal Wojciechowski	Poznan University of Ecomics, Poland
Qiang Zhu	University of Michigan, Dearborn, USA

KMIS Additional Reviewers

Ricardo Fernandes	Centro de Desenvolvimento de Sistemas, Brazil
José Guilherme Mayworm	Federal University of Rio de Janeiro, Brazil

Invited Speakers

Dieter A. Fensel	University Innsbruck, Austria
Frans Coenen	University of Liverpool, UK
Una-May O'Reilly	MIT Computer Science and Artificial Intelligence Laboratory, USA
Marijn Janssen	Delft University of Technology, The Netherlands
Daniel Chirtes	Executive Agency for SMEs - European Commission, Romania

Contents

Knowledge Discovery and Information Retrieval

Knowledge Engineering and Ontology Development

Knowledge Management and Information Sharing

Knowledge Discovery and Information Retrieval

Image Representation for Image Mining: A Study Focusing on Mining Satellite Images for Census Data Collection

Frans Coenen[1(✉)] and Kwankamon Dittakan[2]

[1] Department of Computer Science, The University of Liverpool,
Liverpool L69 3BX, UK
coenen@liverpool.ac.uk

[2] Faculty of Technology and Environment, Prince of Songkla University (PSU),
Phuket, Thailand
kwankamon.d@phuket.psu.ac.th

Abstract. This paper firstly presents a taxonomy for mage representation in the context of image mining. The main premise being that the actual mining algorithms that may be used are well understood, it is the preprocessing of the image data that remains a challenge. The requirement for the output from this preprocessing is some image representation that us both sufficiently expressive while at the same time being compatible with the mining process to be applied. Three categories of representation are considered: (i) statistics-based, (ii) tree-based and (iii) point series based. The second contribution of this paper is an analysis of the proposed representations categories with respect to a novel image mining application, the collection of individual household census data from satellite imagery, more specifically Google earth satellite imagery. The representations are considered both in terms of generating census prediction models and in terms of applying such models for larger scale census prediction.

1 Introduction

Image mining is an important element of the canon of data mining. Decision making is routinely supported by visual information and visualisations of data. At the same time our ability to collect visual information (image data) is increasing rapidly, partly because of technological advancements and partly (and associated with the first) the increasingly reduced cost of collecting such data. For example the collection of retina images are now routine for anyone visiting an optician, whilst the cost of Magnetic Resonance Imaging (MRI) scans has reduced considerably. The computing power available to process images is also rapidly increasing. Consequently the demand for utilising image data for the purpose of extracting information (image mining) is increasing. It should also be noted here that the images we wish to process, although typically 2D in nature, can also be in a 3D format; our ability to collect 3D (volumetric) data has also been advancing such that 3D data is now also readily available.

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 3–27, 2019.
https://doi.org/10.1007/978-3-319-99701-8_1

The challenge of image mining is not so much the algorithms used to extract knowledge from image data, these tend to be well understood, but the end to end process from the initial image representation to the final knowledge interpretation. Although our ability to process large quantities of data is increasing, typically we are still not able to represent image data in its entirety (pixel by pixel), nor in most cases would this be useful; although the use of techniques such as the Convolutional Neural Networks (CNNs) [23] is a significant step in this direction. The manner in which we represent the image data we wish to mine is of great significance (the "rubbish in rubbish out" aphorism is applicable here). This paper seeks firstly to provide an overview of image representation for image mining by considering an "image representation for image mining" taxonomy. Secondly this paper seeks to present this taxonomy in the context of a particular application domain, the mining of satellite imagery to collect census data.

Broadly, image representation for image mining can be viewed according to whether we wish to consider an entire image or simply one or more elements within an image. The later requires the representation process to be preceeded by a segmentation process so as to isolate the elements of interest (segmentation is outside the scope of this paper). In both cases similar techniques can be used for representation purposes, the distinction is the amount of storage that might be required. The most common representation used for data mining in all its forms is the feature vector representation where we conceptualise the domain of interest in terms of an n dimensional feature space where each dimension is an attribute contained in the domain of interest. Using the feature space concept each example (record or image) is defined in terms of a *feature vector* $V = \{v_1, v_2, \ldots, v_n\}$ where each element v_i relates to an attribute value in dimension i. Thus, using the feature vector mechanism, prior to the application of any mining activity, it is necessary to first preprocess the data so that a collection of feature vectors, $\varPhi = \{v_1, V_2, \ldots, V_m\}$, can be generated. The image mining domain is no exception, the challenge is identifying the image features to be included in the feature space. In this context three categories of representation are considered in this paper: (i) statistical, (ii) graph based and (iii) point series. A second challenge is that frequently the number of dimensions is large (the "curse of dimensionality" aphorism is also applicable here). Many of these dimensions (features) are likely to redundant or not useful. To reduce the set of dimensions a feature selection process is typically adopted. Although not a central theme of this paper a number of feature selection methods (χ^2, Gain ratio and Information gain) were considered respect to the population estimation application domain.

There are a great variety of data mining techniques applicable to image data and data in general [12]. So as to limit the scope of the work presented in this paper the focus is on prediction (classification) using supervised learning. A process whereby prelabelled training data is required from which a predictor can be "learnt". Thus the desired feature vectors used for training purposes need to include a class attribute value c drawn from a set of such values C. Thus in this case the feature vectors are of the form $\{v_1, v_2, \ldots, v_n, c_i\}$ where $c_i \in C$. To obtain some degree of confidence in a generated predictor a further prelabelled test set is required to which the prediction model can be applied and the generated predictions (classes) compared to the known predictions. Once we are satisfied with the operation of our predictor it can go into usage and be used to label previously unseen data.

To illustrate the ideas concerning image representation presented in this paper a census collection application domain is considered. A census is a mechanism for acquiring and collecting information about a population; a mechanism widely used with respect to a variety of national, and local, government management and planning activities. The most important element of a census is population count. However, census collection, and the associated post processing, is expensive. The UK Office for National Statistics (UKONS) reported that the UK 2011 census cost some £480 million [33]. The US 2010 census was reported to have cost $13 billion [30]. The cost of census collection is also increasing, according to the Australian Bureau of Statistics the Australian 2006 census cost around AUD 300 million, whilst the 2011 census cost around AUD 440 million [35]. The cost with respect to rural areas is typically greater than that of urban areas because the communication and transport infrastructure in rural areas tends to be less well developed. There is also often a lack of good will on behalf of a population to participate in census collection, even if they are legally obliged to do so, because people are often suspicious of the motivation behind censuses.

A potential solution is the usage of technology, namely the internet. However, this requires a literate population and access to the necessary infrastructure. In many parts of the world people remain unconnected to the internet. It is also interesting to note that in the context of the UK 2011 census it was found that a frequently cited reason for households not to have internet access was because of a "life style" decision not to do so [29]. The solution advocated in this paper is to create a prelabelled training set of household images, extracted from Google Earth, with known family sizes and use this data to build a household size prediction model that can then be used for large scale census collection exercises. Of course it is acknowledged that this approach will not work well in cities where it will be difficult to distinguish buildings in terms of number of inhabitants, however, it was anticipated that the approach would work well in rural areas; areas where census data collection tends to be more of a challenge. In the context of this census collection application domain, and with reference to the proposed image representation for image mining taxonomy, the challenge is how best to represent the satellite household image data. A number of different representations (using the proposed taxonomy) are considered in this paper, evaluated and utilised with respect to a large-scale study featuring a rural area of Ethiopia.

The rest of this paper is structured as follows. In Sect. 2 a review is presented, founded on work presented in [7], regarding existing work on automating the census collection process using satellite image data. The proposed image representation for image mining taxonomy is then presented in Sect. 3. The proposed solution to the automated extraction of census data from satellite imagery, using the proposed taxonomy presented in the previous section, is then presented in 4. Three different catagories of representation are considered and evaluated as presented in Sect. 5. The main findings are presented in the concluding section, Sect. 6.

2 Previous Work

In this section some discussion concerning previous work on population estimation is presented; the application domain focus for the discussion on image representation for data mining presented later in this paper. Population estimation has been a subject of researched amongst the Geographic Information Systems (GIS) and remote sensing communities for some time. From the literature we can broadly divide this research activity into two categories: (i) area interpolation and (ii) statistical modelling [42]. The work presented in this paper subscribes to the second. Using area interpolation the idea is to use existing census information concerning some geographic area and extrapolate this to obtain a population estimation for a wider or alternative geographic area [24]. Statistical modelling in turn is concerned with the relationship between population size or density and data obtained from GIS and/or satellite imagery.

Existing work on statistical modelling for the purpose of population estimation can be further categorised according to the nature of the data on which the population estimation generation is based, namely: (i) light intensity, (ii) land usage, (iii) dwelling unit count, (iv) image pixel characteristics and (v) physical or socio-economic characteristics. The central idea on which the first is based is that there is a functional relationship between population size and the amount of night time light emanating from an area. Examples can be found in [3,6,28,36,39], where the relationship between population density and light frequency were used to convert light frequency into a population density estimation. In [36] the reported evaluation was directed at Japan and China, whereas in [6] and [28] it was directed at China only; in [3] the evaluation was directed at a population estimation of the Brazilian Amazon, whilst in [39] the study was directed at the USA.

Work within the second sub-category is directed at the correlation between population density and different types of land usage. The idea is to determine population density according to land usage with respect to a set of one or more sample areas and apply this knowledge to additional areas (as in the case of area interpolation). Land usage categories are typically identified from satellite image data. In [22] it is suggested that population densities for different types of land usage can be determined from sample surveys or census statistics. Four different types of land usage were extracted from four different cities in California, USA, and population densities computed. In [27], six types of land usage were identified in the context of Landsat TM satellite images centred on Atlanta, USA. A regression model was then applied to produce population densities for Atlanta. A common way of distinguishing land usage types is by using some form of texture analysis. There are a variety pf texture analysis techniques that can be adopted, but one involves the usage of Local Binary Patterns (LBPs) [25,34], a techniques utilised with respect to the work presented in this paper (see Sect. 4).

The third sub-category of approach to population size estimation using statistical modelling is to estimate the total "dwelling unit" count in a defined region and multiply this by an average number of people expected to live in a dwelling unit. There are various ways of obtaining an estimate of the dwelling unit count, but one suggested approach is to estimate this by analysing remote sensing images; a idea also promoted in this paper. In the past, when there was no effective ways of automatically identifying residential buildings within remote sensing imagery, the dwelling units were manually identified from aerial photographs (a laborious and time consuming process). With the advancement of technology and the availability of satellite imagery more advanced "feature extraction techniques" have been developed for this purpose [13]. In [4] a dwelling unit count based approach is presented using IKONOS satellite images of the Al Shaabia district in Khartoum, Sudan. The dwelling unit count approach has some similarity with respect to the work represented in this paper.

In the fourth sub-category, the relationship between image pixel characteristics and population densities is used for the purpose of population estimation. The image pixel characteristics in question can be represented using a variety of mechanisms, but common examples include mechanisms based on the spectral reflectance values of image pixels and mechanisms based on image texture analysis. Examples of using pixel characteristics for population estimation are presented in [18] and [24]. In [18] a system was presented whereby texture analysis was applied to Google Earth satellite images, using block sizes of 64×64 and 32×32 pixels, to estimate population densities with respect to cities in Pakistan. In [24] a variety of features were used, including: spectra signatures, principle components, vegetation indices, fraction images, texture and temperature. These features were extracted from Landsat ETM+ satellite images and used to measure population density in the city of Indianapolis, Indiana, USA.

The final category of population estimation is founded on the usage of various kinds of physical and socioeconomic information which is then interpolated to give population estimations. For example, information about demography, topography and transportation networks have all been used to estimate population size. In [26] a mechanism was presented for estimating population size by determining the correlation between the population in urban areas and the distance to the nearest Central Business Distract (CBD), distance to major roads, slope and the age of the community.

What all the above approaches to population estimation modelling have in common is that they are focussed on regions or areas rather than specific households as in the case of the work presented in this paper. As far as the authors are aware the fundamental approach of estimating populations sizes at the household level, as presented in this paper, is entirely unique.

3 The Image Representation Taxonomy

The proposed image representation for image mining taxonomy is presented in Fig. 1. From the figure, at a high level and as noted in the introduction to this paper, the image representations can be categorised according to whether we are interested in entire images or some region (object) within an image set. We use the terms *global* and *local* to differentiate between the two. An example where we might wish to consider entire images, as shown in the figure, is in the case

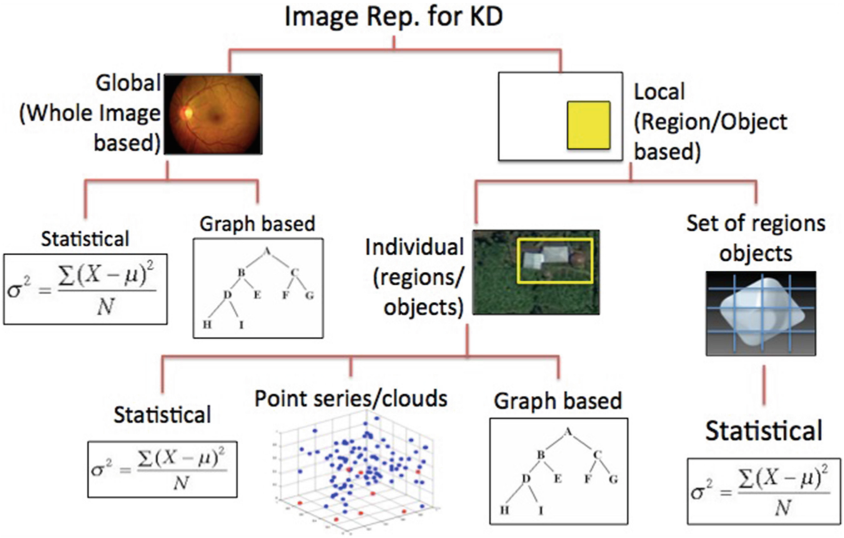


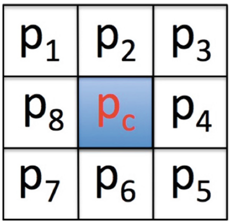
Fig. 1. Taxonomy for image representation for image mining.

of retina image analysis where we typically wish to classify images according to whether they feature some eye condition or not. The most common eye conditions that are considered in this respect are Age-related Macular Degeneration [2, 15, 45] and Diabetic Retinopathy [1, 37]. Frequently quoted examples where we might wish to consider objects within images are with respect to MRI brain scan data; in 2D the mid-sagital slice is often used. For example in [9] the object of interest was the corpus callosum, the part of the brain that connects the left and right hand sides of the brain, which was analysed in the context of the presence of epilepsy (or not). In [41] the ventricles are considered but in terms of 3D image analysis.

In Fig. 1 a distinction is made between non-contiguous region/object applications (such as the census collection based on individual household sizes domain considered later in this paper) and contiguous region/object applications where we have multiple neighbouring regions/objects. In the figure the latter is illustrated with a sheet metal forming 3D image application taken from [21] (see also [11]) where a pre-specified shape is "pressed" out using a sheet metal forming machine. However, the process introduces distortions (referred to as *springback*). The idea presented in [21] was that if these distortions can be predicted they can be compensated for. Sub-shapes in the manufactured shape were thus isolated and considered to be objects in a 3D image mining exercise. The feature vectors in this case comprised a shape description, not unlike the shape descriptions used in [41] to represent ventricles extracted from MRI scans, and a numeric distortion class label. Where we have multiple disconnected objects in an image these can be processed in the same way as if there were only one object.

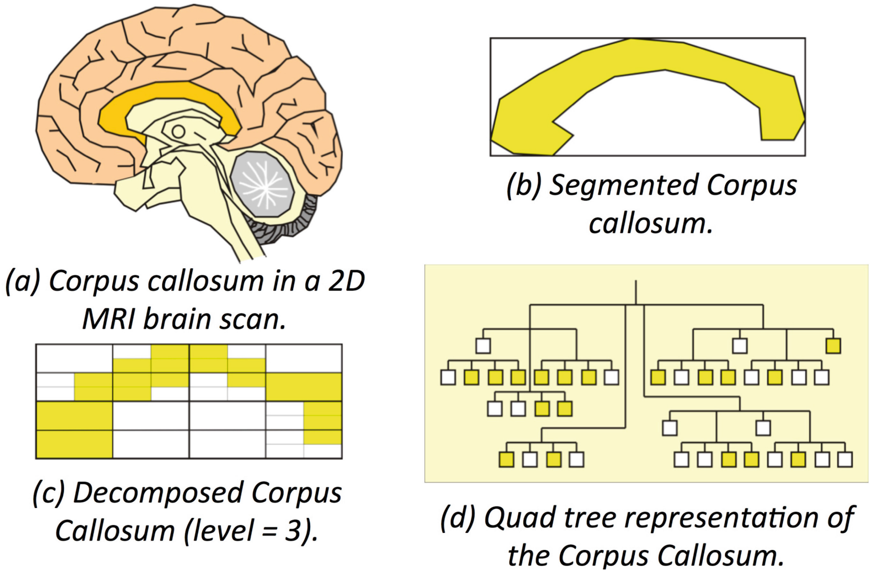
Regardless of whether the nature of images are considered globally or locally they can be represented either in terms of a set statistics extracted from the image (object) or in terms of some graph/tree representation. In the case of single objects we can also consider the boundary of the object which can then be represented as a point series or curve. Statistical techniques are the simplest approach to image representation for image mining. The most obvious statistics that may be used are the first order statistical functions such as the mean, variance and standard deviation of the intensity or RGB or grayscale colour values. For example in [44] (see also [43]) eleven different first order statistical features were extracted from a breast biopsy image set in order to predict the presence of breast cancer (or otherwise). In the case of objects we can use morphometrics of various kinds describing the size and/or shape of an object (size can be expressed simply in terms of a pixel count). Such simple statistics often do not work well because they are not expressive enough; however they provide for a good bench mark representation and are considered later in this paper as a means of representing households. A more sophisticated category of statistic involves the usage of second order statistical functions applied to an intermediate representation, examples include: (i) co-occurrence matrices, (ii) gradient analysis, (iii) Hough transforms and (iv) Local Binary Patterns (LBPs). This last is used later in this paper and thus is considered in further detail later in this section. Another example where the LBP concept has been used as an image representation for image mining can be found in [8] where X-ray images of knee joints are encapsulated using LBPs for the purpose of predicting the existence of osteoarthritis (or otherwise).

A LBP is a texture representation method which is statistical in nature [25, 34]. Using the LBP approach a binary number is produced for each pixel, by thresholding its value with its neighbouring pixels. Thus, with reference to Fig. 2 the grayscale value for the centre pixel p_c is compared with that of the eight neighbours and a value of 1 recorded if the value for p_i is lower than that for p_c, and a value of 0 otherwise. In this manner we get eight binary values making up an eight bit number. There are 256 different options, thus we can generate a 256 dimensional feature space with each dimension having values of between 0 and the maximum number pixels that can exist in any one image in the image set. The example in Fig. 2 considers eight neighbouring pixels at a radius of one, thus 8 × 1 LBPs. There are other possibilities, for example 8 × 2 or 16 × 2.



$$d = \begin{cases} 1, & \text{if } p_i \leq p_c \\ 0, & \text{otherwise} \end{cases}$$

Fig. 2. Ilustration of the LBP concept.


(a) Corpus callosum in a 2D MRI brain scan.

(b) Segmented Corpus callosum.

(c) Decomposed Corpus Callosum (level = 3).

(d) Quad tree representation of the Corpus Callosum.

Fig. 3. Hierarchical decomposition, an example using MRI brain scan data.

A alternative popular method for representing images is to apply some form of hierarchical decomposition to the image (with respect to both the global and local situations) and to store the result in a quad-tree (for 2D image data) or oct-tree (for 3D image data). Hierarchical decomposition has a well established track record in the context of image analysis [31,38,40]?. An example decomposition is given in Fig. 3, based on [10], where the Corpus Callosum featured in a 2D MRI brain scan image has been segmented and decomposed (down to a maximum decomposition level of 3) and rendered as a quad tree. Once we have a collection of tree represented images/objects we can apply a subgraph mining technique (a good review of such techniques is given in [19]) to the tree set and extract frequently occurring sub-trees where frequent is defined in terms of some frequency count threshold σ. A popular frequent subgraph mining algorithm used for this purpose is the gSpan algorithm [17] adapted for the purpose of frequent sub-tree mining rather than frequent sub-graph mining. The set of extracted frequent subgraphs can then be viewed as features in a n-dimensional binary-valued feature space where n is the number of sub graphs and each dimension has two values: present and not present. Issues with hierarchical decomposition include: (i) the "boundary problem" where regions appear in different branches of the tree and (ii) when to stop the decomposition (using either a critical function to measure homogeneity or a pre-specified maximum level of decomposition).

Although quad trees are the most commonly encounter tree-based formalism other types of tree (and graph) format can be adopted. In Fig. 4 (taken from [14]) an alternative decomposition is shown with respect to a retina image. In this case the decomposition alternates between "angular" and "circular" division.

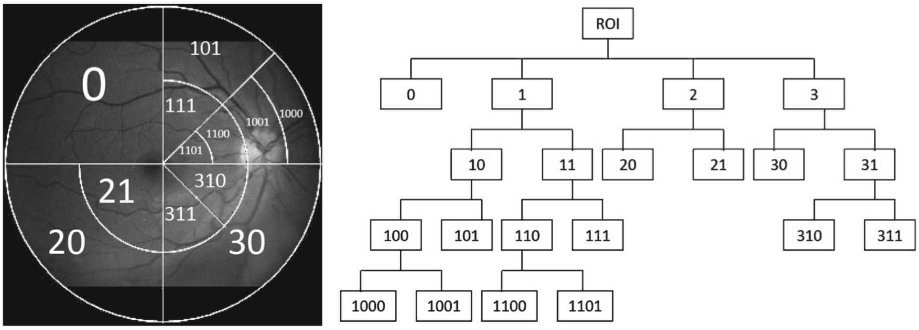


Fig. 4. Circular whole image decomposition with respect to a retina image (max level of decomposition = 4) and Associated tree structure [14].

Angular division involves partitioning using a minor arc to divide a region into two, whilst circular decomposition involves dividing a region into two using a radius emanating from the centre of the image. At the top level the image is divided into four quadrants; at subsequent levels the decomposition is conducted in a binary manner as indicated by the example tree shown on the right of the image.

Another popular mechanism for representing images is as a point series (or curve). Although not indicated in Fig. 1 this can also be applied globally. The simplest form of point series is a histogram, which can be directly translated into a feature vector representation. For example histograms of intensity values, orientation gradients or LBPs. Alternatively, given an object of interest contained within an image, we can represent the boundary in terms of a point series using, for example, the concept of chain coding. Given a collection of labelled point series, representing a set of images or objects within an image set, a new image/object can be classified directly using (say) the well-established KNN algorithm ($k = 1$ is often used). When using algorithms such as KNN we need an appropriate similarity measure; Euclidean distance is frequently used as a comparison measure but requires the point series to be of the same length. Alternatively, we can look to work on time series analysis [20], for example the use of Dynamic Time Warping (DTW) which produces a "warping path distance" defining the difference between two point series [5,32]. DTW has the added advantage that the point series to be compared do not have to be of the same length. The later was used in [9] to define shapes in 2D MRI brain scan data and in [41] to define 3D MRI brain scan shapes.

4 Census Prediction Model Generation

In this section we return to the census collection estimation application domain. Recall that the idea is to build a household size predictor using prelabelled household images extracted from Google Earth and then to use this predictor

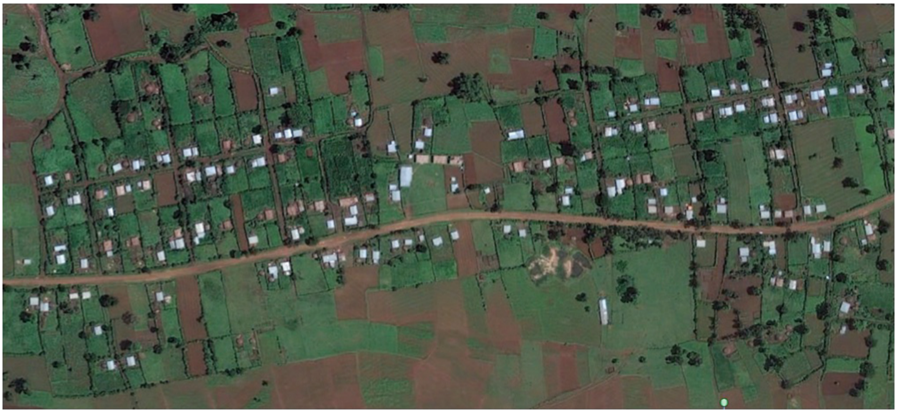


Fig. 5. Google Earth image featuring a village in the Horro district of Ethiopia [7].

to produce population estimations of large areas. Recall also that to act as a focus for this work a rural area of Ethiopia was selected. More specifically the district of Horro located 300 Km to the northwest of Addis Ababa. An example Google earth image from this area is presented in Fig. 5. Inspection of the image indicates a large number of households.

To collect the required training data an "on the ground" team visited sample households at two sites, Site A and Site B, within the district and collected family size information together with the latitude and longitude of each household so that the associated Google satellite images could be retrieved. At the time the data was collected Google Earth did not readily facilitate the automated extraction of satellite imagery, so instead the Google Static Map Service was used. This featured an API that allowed users to download satellite images (one image at a time) specified according to various parameter settings: (i) latitude and longitude of the centre of the area of interest. (ii) image size (in pixels) and (iii) zoom Level (level of detail). An image size of 1280 × 1280 pixels and a zoom level of 18 was used. Each household was surrounded with a 256 × 256 pixel bounding box defined so as to cover the largest anticipated household (by superimposing a box we do not have issues with irregular shaped household plots). In this manner data for 120 households was obtained, 70 households for Site A and 50 for Site B. The distinction between the two sites was that for Site A the available Google Earth images were obtained in the "wet season" and so were mostly green, while those obtained for Site B were obtained during the dry season so were mostly brown. Some statistics concerning this training data are given in Table 1 and Fig. 6. Note from the table that the population sizes have been grouped according to three class labels: (i) Small, (ii) Medium and (iii) Large. The reason for this was for prediction purposes categorical classification systems as well as regression models (which produced a real value) were considered; as described in further detail later in this section.

Table 1. Statistics for training data (Sites A and B).

Family	Min.	Max.	Ave.	Mode	Site	Site
Small	2	5	4.04	5	38	19
Medium	6	8	7.00	6	32	21
Large	9	12	9.80	9	10	10
All	2	12	6.31	6	70	50

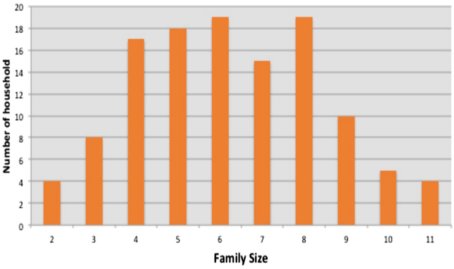


Fig. 6. Histogram for training data household population sizes (Sites A and B) [7].

Given this training data the next stage was to represent the households using an appropriate mechanism compatible with prediction model generation. To this end each of the three categories of representation identified in the taxonomy presented in Sect. 3 was used and a comparison conducted. Details concerning each individual representation are given in Sub-sect. 4.1, 4.2 and 4.3 below; and the conducted comparative evaluation reported on in Sub-sect. 4.4.

4.1 Satistics Based Image Representation

For the statistics based representation LBPs were used, generated as described in Sect. 3. LBPs with eight neighbours and a radius of one were used (8×1 LBPs). Experiments were conducted (not reported here) using other LBP configurations but no advantage was found. An example of the process of converting a Google household image to an LBP image is given in Fig. 7. The left hand image shows the raw Google image, the middle image the associated grayscale image to which the LBP mechanism was applied and the right hand image the resulting LBP rendition of the original Google image.

Once counts for each of the 256 possible LBPs had been obtained a 256 element feature vector could be generated, one for each household image. Feature selection was then applied so as to reduce the overall number of dimensions and retain dimensions which were good discriminators of household size. A subset of the LBPs was thus retained. A number of feature selection methods were considered but χ^2 feature selection, with $k = 40$ (where k is the number of dimensions to be retained), was found to produce the best result. Consideration was also given to augmenting the LBP representation with additional statistics (such as contrast, correlation, energy, homogeneity) but this was also found to have little effect.

4.2 Tree Based Image Representation

The tree-based image representation generation process is illustrated in Fig. 8 (see also Fig. 3) where we start with a colour Google image (not shown), convert this to a grayscale household image and the apply the decomposition. In

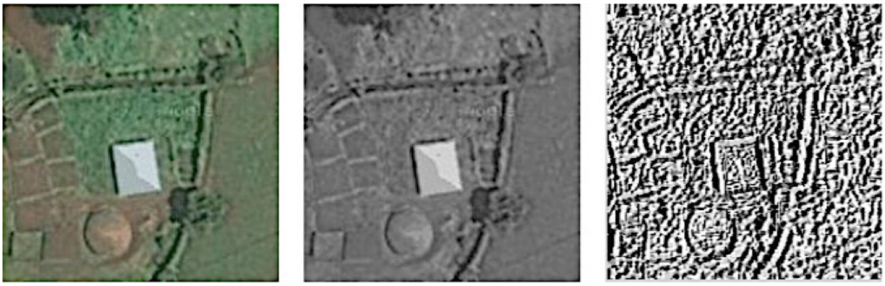


Fig. 7. Process of converting a Google household image to an LBP image [7].

this manner a quadtree was generated representing each household in the training data. The nodes were labelled with a greyscale encoding generated using a mean intensity of the greyscale colours in each region; for this purpose eight labels were derived, each describing a range of 32 consecutive intensity values. Frequent Sub-graph Mining (FSM) was then applied to the tree collection as discussed in Sect. 3. A variation of gSpan was used, but other FSM algorithms would be equally applicable. A σ value of 10 was used for the FSM; in other words for a sub-graph to be considered frequent it had to appear in 10% of the tree represented images in the training set. Note that low σ values are better (nothing will be missed), however many more Frequent Sub-Graphs (FSGs) will be identified than when a higher σ value is used. A feature selection strategy was thus adopted so as to reduce the number of dimensions in a manner whereby only highly discriminative features were retained. In this case gain ratio feature selection, with $k = 55$, was found to produce the best result. Each record was then presented in a feature vector format ready for prediction model generation.

4.3 Point Series Based Image Representation

For the point series based household mage representation seven different colour histograms were generate: three describing the three channels in the RGB image colour formalisation, three describing the three channels in the HSV image formalisation, and one using the greyscale formalisation. For each histogram 32 bins were used, thus feature vectors measuring $7 \times 32 = 224$ elements were generated. The authors again experimented with including statistical measures, but it was again found that this made no difference (and in some cases proved to be decremental). Feature selection was also applied to reduced the size of the feature space. As before a number feature selection strategies were considered (χ^2, Gain ratio and Information gain). Gain ration Feature Selection with $k = 25$ was found to give the best results.

4.4 Household Image Representation Evaluation

This sub-section presents a comparison of the above suggested image representation techniques, one for each representation category included in the proposed

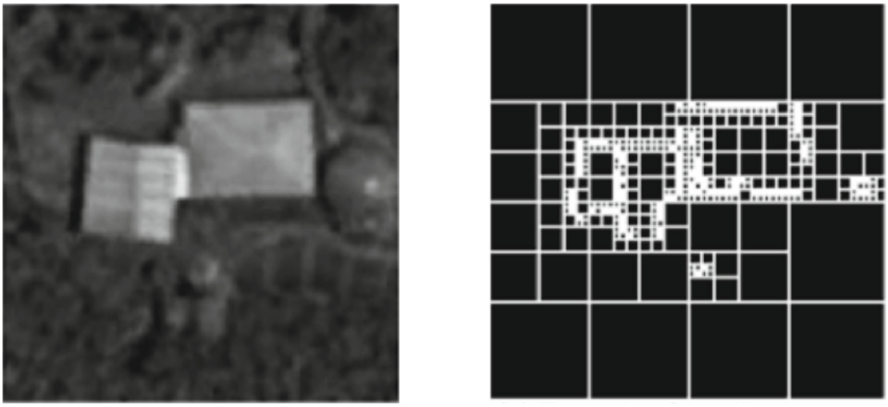


Fig. 8. Hierarchical decomposition process for an example Google household image [7].

image representation for image mining taxonomy. The comparison was conducted by generating and testing predictors using each representation. For this purpose two different categories of prediction model were considered: (i) classification models and (ii) regression models; the distinction being that the first is used to predict a class label while the second is used to predict a real value.

In the classification case three different classes were considered (as given in Table 1) and six classifier generation models: (i) the Bayesian Network (BN) model, (ii) the Neural Network (NN) model, (iii) Logistic Regression (LR), (iv) Sequential Minimal Optimisation (SMO), (v) Averaged One Dependence Estimation (AODE) and (vi) the well known C4.5 decision tree generation algorithm (C4.5). These were coupled with χ^2 and Gain ratio feature selection (with different values of k). The metrics used for the evaluation were: (i) Accuracy (AC), (ii) Area Under receiver operating Characteristic (AUC), (iii) the F-measure (FM), (iv) Sensitivity (SN) and (v) Specificity (SP). The best results, generated using Ten Cross Validation (TCV), are given in Table 2 (derived from work included in [7]), with the very best results highlighted in bold font. For the statistics-based household image representation χ^2 feature selection with $k = 40$ was used. For the tree-based image representation $\sigma = 10$ was used for FSG mining and gain ration feature selection with $k = 55$. For the point series representation gain ration feature selection was also used but with $k = 25$. From the table it is interesting to note that the "best" representation depends on whether we have wet (green) season or dry (brown) season image data. For the wet season the statistics based representation, using LBPs and coupled with either LR or NN, provided best results; while for the dry season the tree based representation (coupled with BN) provided the best results.

In the context of the linear regression models a number of models were considered: (i) Linear regression (Linear Reg.), (ii) Least Median Squared regression (LMedS), (iii) Isotonic Regression (IsoReg) and (iv) Support Vector Machine regression (SVMreg). Each was considered in isolation and when coupled with different feature selection strategies. Each was applied in the context of the satellite household image training data expressed using the different representations

Table 2. Household image representation evaluation using classification models and TCV.

Classification model generator	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
Stats. based (LR)	**0.771**	0.859	**0.778**	**0.771**	**0.885**	0.680	0.756	0.679	0.680	0.803
Stats. based (NN)	**0.771**	**0.881**	0.759	**0.771**	0.852	0.720	0.824	0.718	0.720	0.825
Tree Based (AODE)	0.629	0.815	0.627	0.629	0.753	0.800	0.863	0.785	0.800	0.871
Tree based (BN)	0.600	0.808	0.596	0.600	0.734	**0.800**	**0.879**	**0.792**	**0.800**	**0.876**
Tree based (NN)	0.686	0.819	0.685	0.686	0.782	0.620	0.789	0.628	0.620	0.829
Tree based (SMO)	0.729	0.791	0.727	0.729	0.818	0.620	0.733	0.610	0.620	0.781
Point based (BN)	0.700	0.807	0.687	0.700	0.782	0.700	0.798	0.692	0.700	0.829
Point series (C4.5)	0.671	0.724	0.668	0.671	0.760	0.500	0.598	0.499	0.500	0.718
Point based (LR)	0.657	0.822	0.662	0.657	0.806	0.640	0.821	0.633	0.640	0.798

discussed above. The metrics used for comparison purposes in this case were: Correlation Coefficient (Coef), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Note that metrics used for evaluating classification models, where we wish to predict a categorical class, tend to be different to those that are typically used for evaluating regression models where we are predicting a real value. Best results, again generated using TCV, are presented in Table 3 (also derived from work included in [7]), with the very best results again highlighted in bold font. In this case the LBP statics-based image representation outperformed the other representations hence only results using the LBP representation are shown. The table also gives results with and without the application of a feature selection strategy. A number of such strategies were considered, but Correlation-based Feature Selection (CFS) was fund to produce the best results. From the table it can be seen that the best performing regression model was SVM regression.

Table 3. Household image representation evaluation using regression models and TCV.

Regression method	Site A			Site B		
	Coef	MAE	RMSE	Coef	MAE	RMSE
LinearReg	−0.080	2.167	2.570	0.274	1.981	2.407
LMedS	−0.288	3.262	3.894	0.215	1.952	2.353
IsoReg	−0.309	2.382	2.841	0.156	1.940	2.295
SVMreg	−0.279	3.367	3.970	0.308	1.778	2.056
LinearReg+CFS	0.084	2.145	2.550	0.400	1.727	2.093
LMedS+CFS	0.252	1.988	2.373	0.428	1.687	2.038
IsoReg+CFS	−0.202	2.287	2.706	0.109	1.912	2.282
SVMreg+CFS	**0.307**	**1.957**	**2.330**	**0.587**	**0.143**	**1.802**

5 Large Scale Study

Once a prediction model has been generated and tested so that an appropriate degree of confidence can be attached to the model it can be placed into service. This section considers firstly how the models, generated as described in the previous section, can be applied in the context of regional census collection. The section firstly presents the process whereby such a census might be conducted and secondly considers the effectiveness of the result by considering a particular benchmark region. The test area chosen for this purpose was an entire village and its surrounding lands within the Horro district. The reasons why this area was chosen was because this area was similar to the areas from which the prediction model training data was obtained and because the population size of this village was known; in 2011 the village was reported to comprise 459 households and a population of 3,223 (thus ground truth data was available).

The rest of this section is organised as follows. Sub-sect. 5.1 describes the satellite data collection process using the Google Static Map Service API. Once the satellite image data has been collected the images need to be segmented to identify "household images"; the mechanism whereby this was conducted is presented in Sub-sect. 5.2. To ensure no data was missed an overlap was used when collecting the satellite image data, thus it was possible that specific households would appear in more than one image. It was thus necessary to first remove such duplicates before any further processing could be conducted. The duplicate household detection and pruning process is considered in Sub-sect. 5.3. Once the household images had been identified they could be represented using one of the image representations considered above, to which any of the prediction models also considered above could be applied. To evaluate the process the best performing prediction models (see Sub-sect. 4.4) were applied to the data. The results are presented and discussed in Sub-sect. 5.4.

5.1 Satellite Image Data Collection

In total 600 Satellite images, covering the area of interest were collected using the Google Static Map Service API. An image size of 1280 × 1280 pixels and a zoom level of 18 was used; because these were the parameters used for the training set collection. Using the Google Static Map Service API images are downloaded in an iterative manner image by image. A 320 pixel overlap was used, designed so that every household will appear in its entirety in at least one collected satellite image. For this to operate correctly it was necessary to: (i) convert the top-left corner latitude and longitude of the current image into x and y pixel values, (ii) add the required offset to obtain the top-left x and y coordinates of the next image in the sequence, (iii) convert these new x and y coordinates back to a latitude and longitude and (iv) repeat; a time-consuming process. Note that Cartesian coordinates are planer values while latitude and longitude are geoidal values, thus conversion was also not straight forward; note that the Google Static Map Service uses the EGM96 spheroid (Earth Gravitational Model 1996). It took 356 s to collect the 600 required satellite images. Together these images formed

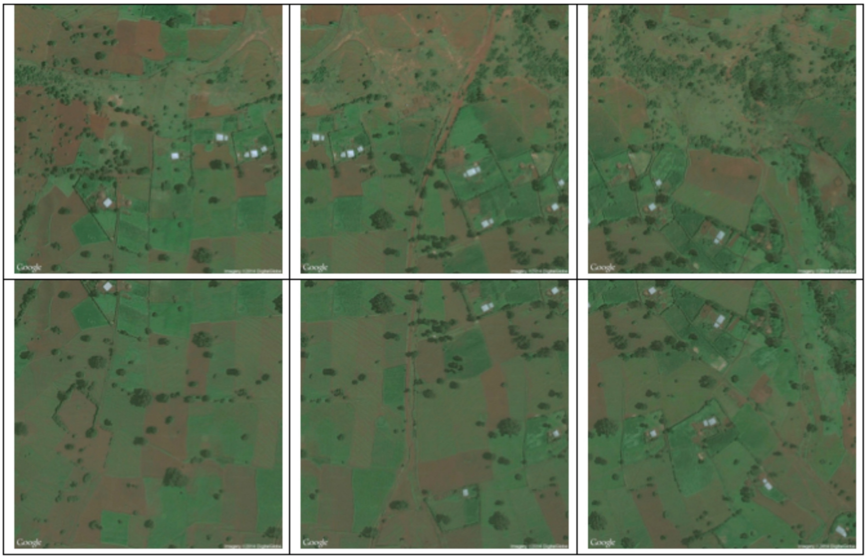


Fig. 9. Fragment of collected satellite image patchwork [7].

a "patchwork" covering the area of interest. A fragment of this patchwork is shown in Fig. 9.

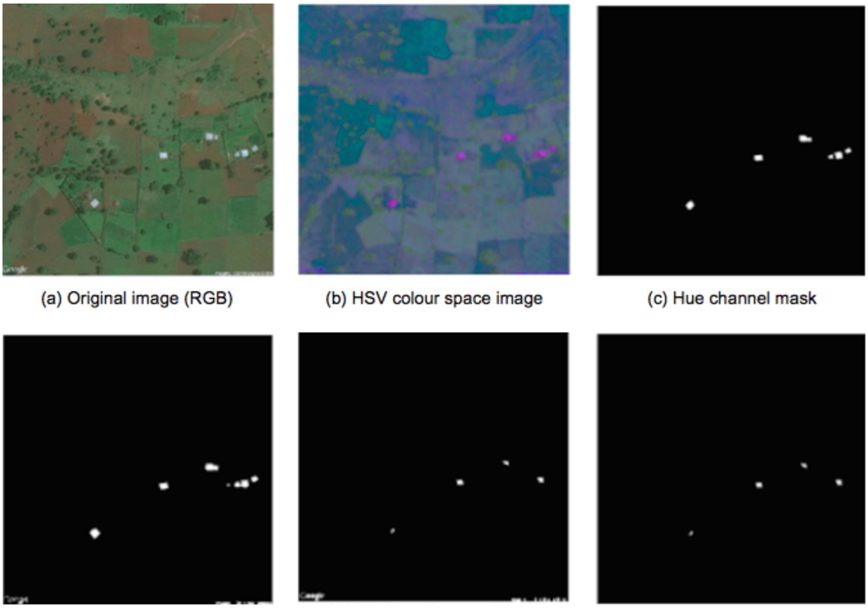
5.2 Image Segmentation

The downloaded satellite images could contain zero, one or more households. It was thus necessary to segment the images so as to identify households. As noted previously, the typical household comprised at least one building with a tin roof that was readily discernable (see Fig. 5). Visual inspection of a sample of the images indicated that this was true in all the cases sampled. This feature could therefore be usefully employed to identify households in the collected satellite image data. Note that with respect to the training data (Sect. 4) segmentation was not required because we knew where the households were because their latitude and longitude had been collected as part of the knowledge acquisition process.

The segmentation was conducted using a number of image masks. Experiments were conducted using a variety masking techniques (a significant challenge was the illumination of roads and water ways). The most appropriate mechanism was found to be when the HSV representation was used together with a set minimum and maximum thresholds. Given an image represented using one of the HSV channels, pixels with values below and above the threshold were set to black and the remaining pixels within the threshold range to white. Thus three masks were produced: (i) hue, (ii) saturation and (iii) value. By combining these masks pixels set to white in all three masks were identified as households. Extensive experimentation (not reported here) was conducted to determine the most

Table 4. Adopted HSV threshold values for household image segmentation.

Channel	Min	Max
Hue	0.35	0.65
Saturation	0.05	0.15
Value	0.80	1.00



(a) Original image (RGB)

(b) HSV colour space image

(c) Hue channel mask

Fig. 10. Illustration of the satellite image segmentation process [7].

appropriate threshold values, the selected values are given in Table 4. The entire segmentation process is illustrated in Fig. 10 where we have an originnl image featuring four households, the image translated into the HSV colour space, three masks (hue, saturation and value) and the final result.

On completion of the segmentation process each household was represented as a "blob" of white pixels. The centroid of each blob was considered to be its location, described in terms of latitude and longitude coordinates, and this location was identified in the original image. Each location in the original image was then surrounded by a $w \times w$ bounding box ($w = 256$ was used as this was the same value used for the prediction model training as described in Sect. 4). In this manner a collection of household images was obtained. Note that the minimum bounding boxes will be smaller and/or non-symmetrical near the edges of each image.

5.3 Duplicate Detection and Pruning

Using the above process 526 household images were identified. However, this included duplicate households; households that appear in more than one image. Inspection of the Fig. 9 indicates a number of duplicate households, some appearing in two images and in some cases in four images. Such duplicate households had thus to be pruned before any further processing could be conducted. The duplicate detection and pruning process was as follows. The identified households were listed in order of latitude. This list was then processed and households with the same latitude and longitude label (within a level of tolerance) identified. If two households with the same centroid latitude and longitude both comprised 256×256 pixel boxes the later was pruned. If the boxes were unequal in size the household featuring the smaller sized box was pruned.

Using the above process a total of 526 households were detected including duplicates. Duplicate detection identified 100 duplicate households, thus 426 out of a "known" number of 459 households were identified. Suggested reasons for the discrepancy were as follows. There was a two year time difference between the "ground truth" survey and the satellite images; a period during which some households may have fallen into disuse (manual inspection of a proportion of the collected satellite images indicated that some buildings did indeed appear to be roofless, thus supporting this conjecture). Inspect of the satellite imagery indicated that a small number of buildings were very poorly defined and in some cases had not been segmented correctly. It was also possible that the duplicate household detection mechanism had detected some duplicates that were in fact not duplicates (although no evidence for this was found). The overall run time to segment and process the collected satellite image data was 1.370 s (22.8 min), about 2.28 s per satellite image and 2.6 s per household.

5.4 Population Estimation Results and Evaluation

Once an appropriate set of household satellite images had been generated previously derived classification and/or regression models (of the form described in the previous section) could be applied and an overall population estimation extracted from the image data. In the case of classification models a class label was produced for each household, to turn this into a population estimation each class needed to be translated into a number of persons and then summed to give a total number of persons. In the case of the classification models described in Sect. 4 the average number of persons associated with each class was used. In the case of the regression models a population size was derived directly.

The obtained results using the best performing classifiers/predictors identified in Sub-sect. 4.4 (see also [7]) are presented in Table 5 (best results highlighted in bold font): (i) Neural Network classifier generated using statistics-based (LBP) Site A wet season data, (ii) Bayesian Network classifier generated using Graph-based Site B dry season data, (iii) SVM Linear regression generated using statistics-based (LBP) Site A wet season data and (iv) SVM Linear regression generated using statistics-based (LBP) Site B dry season data. The

Table 5. Population estimation results.

Representation	Prediction model	Feature selec. strat.	Population estimation	Accuracy (%)	Total run time (Mins.)
Statistics-based (LBPs)	Neural Network classifier generated using Site A wet season data	χ^2	2,545	78.96	29.49
Graph-based ($\sigma = 10$)	Bayesian Network classifier generated using Site B dry season data	Gain Ratio ($k = 55$)	2,495	77.41	35.42
Statistics-based (LBPs)	SVM Linear regression generated using Site A wet season data	CFS	2,548	79.06	**29.48**
Statistics-based (LBPs)	SVM Linear regression generated using Site B dry season data	CFS	**2,760**	**85.63**	**29.48**

Table 6. Estimation of population size with respect to the Neural Network classification model generated using the Site A data set.

Family size	Average household size (a)	Predicted num. households (b)	Estimated population size ($a \times b$)
Small	4.04	156	630
Medium	7.00	261	1827
Large	9.80	9	88
Total		426	2545

Table 7. Estimation of population size with respect to the Bayesian Network classification model generated using of the Site B data set.

Family size	Average household size (a)	Predicted num. households (b)	Estimated population size ($a \times b$)
Small	4.04	226	7913
Medium	7.00	135	945
Large	9.80	65	637
Total		426	2495

calculation of the individual population sizes using Neural Network and Bayesian Network classification is given in Tables 6 and 7 respectively (derived from [7]). The best performing approach used a statistics-based representation coupled with a SVM Linear Regression model. This produced a population estimation of 2,760 compared to a "ground truth" of 3,223; thus an accuracy of 86%.

An accuracy of 86% might be argued to be unsatisfactory, however, we can point to a number of reasons for the difference between the predicted and "ground truth" population sizes. Firstly the data from which the classification (regression) models were generated might not reflect the data to which they were applied as closely as was anticipated. Measures for determining the similarity between satellite image data sets are a subject for future work. Secondly, as already noted, there was a two year time lag between the date of the census collection (2011) and the date of the satellite image acquisition (2013). Manual inspection of a number of images showed signs of derelict (abandoned) households. It may thus be the case that between 2011 and 2013 depopulation had taken place and that the produced population estimates were in fact a better reflection of population size than initially thought. There have been recent reports concerning the depopulation of rural Ethiopia [16]. Thirdly, and again as already noted previously, census collection is often viewed with suspicion. Local authorities may suspect that it is to be used for the levying of a local tax and thus there may be an incentive to under report population size. Alternatively it may be suspected that the census is to be used for allocating development funds in which case there may be an incentive to over report.

6 Conclusions

In this paper we have presented a taxonomy for image representation for image mining together with an illustration of the practical application of the taxonomy in the context of an automated census data collection application. The main premise is that although the data mining algorithms that we might wish to apply to image data are well understood the end to end Knowledge Discovery in Data (KDD) process is less well established. The main challenge is how to represent image data in such a way that the salient features are maintained while at the same time ensuring compatibility with the data mining algorithms to be applied. The proposed taxonomy, at a high level, differentiates between global representations and local representations; the first being directed at applications where we wish to consider image data in its entirety and the second where we wish to consider one or more objects within individual images. A distinction was also made between objects that are connected (contiguous) and not connected. In the taxonomy, again at a high level and regardless of whether we are considering images in their entirety or at a local level, we can identified three categories of representation: (i) statistical, (ii) tree (or graph) -based and (iii) point series based. The particular nature of the individual representations that can be included in the categories depends on whether we are working at a global or local level. At the local level we can, for example, consider the nature of the boundary of the objects of interest which would not be applicable at the global level.

The taxonomy was applied to a census estimation application domain so as to illustrate the usage of the ideas presented by the taxonomy with respect to a novel application domain. The motivation for the application was the resource required to collect census data. The idea was to build predictors to predict individual household sizes from satellite images of households. It was noted that although the idea would not work well in urban areas it would work well in rural areas where the cost of census collection is the greatest. Three exemplar representations were considered, one from each category: (i) a quadtree representation from which frequently occurring sub-trees were extracted and used to generate feature vectors (one per household), (ii) a statistics-based representation founded on the use of LBPs and (iii) a point series representation founded on the use of collections of histograms. Training and test data was obtained from two sites (Site A and Site B) in a rural area of Ethiopia featuring households with a known location and "family size". The known locations were used to obtained Google Satellite images of the individual households. The distinction between the two sites was that for Site A the satellite imagery was obtained during the wet (green) season whilst that for Site B was obtained in the dry (brown) season. The collated individual household images were then represented, using the three selected exemplar representations, to produce three versions of the data. This data was then used to train predictors. Two categories of predictor model were considered. Classification models where a "family size" class label was predicted ("small". "medium" or "large"), and regression models where an actual household size was produced. A range of classification and regression models were considered coupled with different feature selection mechanisms. Of these two classification models and two regression models were found to give the best performance when evaluated using the training/test data and TCV, as follows:

1. Statistics-based using LBPs, Neural Network classification (generated using Site A data) and χ^2 feature selection.
2. Tree-based using Bayesian Network classifier generated (generated using Site B data) and Gain Ratio feature selection.
3. Statistics-based using LBPs, SVM Linear regression (generated using Site A data) and CFS feature selection.
4. Statistics-based using LBPs, SVM Linear regression (generated using Site B data) and CFS feature selection.

The prediction models were then applied to a wider area (but in the same rural region of Ethiopia), an entire village, where the number of households and population size was known. The best performing approach was found to be the LBP Statistics-based representation coupled with a SVM Linear regression model (generated using the Site B data) and CFS feature selection. An accuracy of 85.63% was recorded. Although (at face value) the population estimation produced was not as accurate as the "ground truth" census data (this was to be expected), the proposed method offered significant cost and time savings. A number of reasons as to why the prediction was not identical to the "ground truth" value can be identified:

1. The training data from which the prediction models were generated might not reflect the data to which they were applied as closely as was anticipated. Measures for determining the similarity between satellite image data sets are a subject for future work.

2. There was a two year time lag between the date of the census collection (2011) and the date the satellite images were acquired (September 2013). Manual inspection of a number of images indicated signs of derelict (abandoned) households. It was thus conjectured that it might be the case that between 2011 and 2013 depopulation had occurred and that the produced population estimate was in fact a better reflection of population size than initially thought.
3. Census collection is often viewed with suspicion therefore there may have been incentives to over or under report and therefore the "ground truth" value night not have been entirely accurate (it should not be regarded as a "gold standard").

Whatever the case the results indicated that by using the proposed framework effective population estimates can be obtained, in rural areas, at a very low cost (almost zero).

With respect to image mining in general it can be observed that decisions are regularly made with the support of imagery of some sort (Satellite Image, MRI, OCT, and so on). It can also be observed that our ability to collect imagery of all kinds (both 2D and 3D) has enhanced rapidly over the last decade (we can do it cheaper and faster); we have seen a rapid growth in the global image sensor market. There is substantial benefit to be gained from applying image mining to this image date although it is essential that appropriate image representation is used. There is also a lot of scope for alternative representations, especially fuzzy and deep learning approaches and lots of scope for further application. A further issue to be addressed is explanation mechanisms to give reasons as to why particular predictions and/or classifications were arrived at with respect to previously unseen images; this is of particular relevance with respect to medical imaging applications.

Acknowledgements. The authors wold like to thank the following whose ideas helped formulate the contents of this paper: (i) Abdulrahman Albarrak from the Department of Computer Science at The University of Liverpool, (ii) Ashraf Elsayed from the Department of Computer Science at the University of Alexandria, (iii) Marta García-Fiñana from the Department of Biostatistics at the University of Liverpool, (iv) Hanafi Hijazi from the School of Engineering and Information Technology at the University of Malaysia Sabah, (v) Vanessa Sluming from the School of Health Science at the University of Liverpool, (vi) Akadej Udomchaiporn from King Mongkut's Institute of Technology Ladkrabang and (vii) Yalin Zheng from the department of Eye and Vision Science at the Royal Liverpool University Hospital.

References

1. Albarrak, A., Coenen, F., Zheng, Y.: Classification of volumetric retinal images using overlapping decomposition and tree analysis. In: Proceedings of 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2013), pp. 11–16 (2013)
2. Albarrak, A., Coenen, F., Zheng, Y.: Volumetric image classification using homogeneous decomposition and dictionary learning: a study using retinal optical coherence tomography for detecting age-related macular degeneration. J. Comput. Med. Imaging Graph. **55**, 113–123 (2016)
3. Amaral, S., Monteiro, A.V.M., Câmara, G., Quintanilha, J.A.: DMSP/OLS night time light imagery for urban population estimates in the Brazilian Amazon. Int. J. Remote Sens. **27**(5), 855–870 (2006)
4. Al Salman, A.S., Ali, A.E.: Population estimation from high resolution satellite imagery: a case study from Khartoum. Emir. J. Eng. Res. **16**(1), 63–69 (2011)
5. Berndt, D.j., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Proceedings of AAAI Workshop on Knowledge Discovery in Databases, pp 229–248 (1994)
6. Cheng, L., Zhou, Y., Wang, L., Wang, S., Du, C.: An estimate of the city population in China using DMSP night-time satellite imagery. In: Proceedings of IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS), pp 691–694 (2007)
7. Dittakan, K.: Population estimation mining from satellite imagery. Ph.D. thesis, University of Liverpool (2015)
8. Dittakan, K., Coenen, F.: Early Detection of Osteoarthritis Using Local Binary Patterns: A Study Directed at Human Joint Imagery. In: Booth, R., Zhang, M.-L. (eds.) PRICAI 2016. LNCS (LNAI), vol. 9810, pp. 93–105. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42911-3_8
9. Elsayed, A., Hijazi, M.H.A., Coenen, F., García-Fiñana, M., Sluming, V., Zheng, Y.: Classification of MRI brain scan data using shape criteria. Ann. Br. Mach. Vis. Assoc. (BMVA) **2011**(6), 1–14 (2011)
10. Elsayed, A., Coenen, F., García-Fiñana, M., Sluming, V.: Region of interest based image classification: a study in MRI brain scan categorization. In: Karahoca, A. (ed.) Data Mining Applications in Engineering and Medicine, pp. 225–248. InTech - Open Science, Slavka Krautzeka (2012)
11. El Salhi, S., Coenen, F., Dixon, C., Khan, M.: Predicting springback using 3D surface representation techniques: a case study in sheet metal forming. J. Expert Syst. Appl. **42**(1), 79–93 (2014)
12. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann, Burlington (2011)
13. Haverkamp, D.: Automatic building extraction from IKONOS imagery. In: Proceedings of Annual Conference of the American Society for Photogrammetry and Remote Sensing (2004)
14. Hijazi, M.H.A., Coenen, F., Zheng, Y.: Data mining techniques for the screening of age-related macular degeneration. J. Knowl. Based Syst. **29**, 83–92 (2012)
15. Hijazi, M.H.A., Coenen, F., Zheng, Y.: Data mining for AMD screening: a classification based approach. Int. J. Simul. Syst. Sci. Technol. (IJSSST) **15**(2), 64–68 (2015)
16. Hamza, I.A., Iyela, A.: Land use pattern, climate change, and its implication for food security in Ethiopia: a review. Ethiop. J. Env. Stud. Manag. **5**, 26–31 (2012)

17. Huan, J., Wang, W., Prins, J.: Efficient mining of frequent subgraph in the presence of isomorphism. In: Proceedings of the 2003 International Conference on Data Mining (ICDM 2003), pp. 549–561 (2003)
18. Javed, Y., Khan, M.M., Chanussot, J.: Population density estimation using textons. In: Proceedings of IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS 2012), pp. 2206–2209 (2012)
19. Jiang, C., Coenen, F., Zito, M.: A survey of frequent subgraph mining algorithms. Knowl. Eng. Rev. **28**(1), 75–105 (2013)
20. Karter, J.: Time Series Analysis with MATLAB. CreateSpace Independent Publishing Platform (2016)
21. Khan, M., Coenen, F., Dixon, C., El Salhi, S., Penalva, M., Rivero, A.: An intelligent process model: predicting springback in single point incremental forming. Int. J. Adv. Manuf. Technol. **76**, 2071–2082 (2015)
22. Kraus, S.P., Senger, L.W., Ryerson, J.M.: Estimating population from photographically determined residential land use types. J. Remote Sens. Environ. **3**(1), 35–42 (1974)
23. Krizhevsky, A., Sutskever. I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Proceedings of NIPS (2012)
24. Li, G., Wang, Q.: Using Landsat ETM+ imagery to measure population density in Indianapolis, Indiana, USA. J Photogramm. Eng. Remote Sens. **71**(8), 63–69 (2005)
25. Liang, P., Li, S.F., Qin, J.W.: Multi-resolution local binary patterns for image classification. In: Proceedings of the Twentieth International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), pp. 164–169 (2010)
26. Liu, X., Clarke, K.: Estimation of residential population using high resolution satellite imagery. In: Proceedings of Third International Symposium on Remote Sensing of Urban Area, pp. 153–160 (2002)
27. Lo, C.: Zone-based estimation of population and housing units from satellite-generated land use/land cover maps. In: Mesev, V. (ed.) Remotely Sensed Cities, pp. 157–180. Taylor and Francis, London and New York (2003)
28. Ma, T., Zhou, C., Pei, T., Haynie, S., Fan, J.: Quantitative estimation of urbanization dynamics using time series of DMSP/OLS nighttime light data: a comparative case study from China's cities. J. Remote Sens. Environ. 124, 99–107 (2012)
29. Madden, P., Goodman, J., Green, J., Jenkinson, C.: Growing pains: population and sustainability in the UK. Technical report, Forum for the Future (2010)
30. Mather, M., Pollard, K., Jacobsen, L.A.: Report on America: first results from the 2010 census. Technical report, Population Reference Bureau, Washington, DC, USA (2011)
31. Montanvert, A., Meer, P., Rosenfield, R.: Hierarchical image analysis using irregular tessellations. IEEE Trans. Pattern Anal. Mach. Intell. **13**(4), 307–316 (1991)
32. Myers, C.S., Rabiner, L.R.: A comparative study of several dynamic time-warping algorithms for connected word recognition. Bell Syst. Tech. J. **60**(7), 1389–1409 (1981)
33. Office for National Statistics: National population projections, 2010-based statistical bulletin. Technical report, Office for National Statistics (2011)
34. Pietikäinen, M.: Image analysis with local binary patterns. In: Kalviainen, H., Parkkinen, J., Kaarna, A. (eds.) SCIA 2005. LNCS, vol. 3540, pp. 115–118. Springer, Heidelberg (2005). https://doi.org/10.1007/11499145_13
35. Pink, B.: Census of population and housing: nature and content Australia 2011. Technical report, Australian Bureau of Statistics (2008)

36. Pozzi, F., Small, C., Yetman, G.: Modeling the distribution of human population with night-time satellite imagery and gridded population of the world. In: Proceedings of Future Intelligent Earth Observing Satellites Conference (2002)
37. Pratt, H., Coenen, F., Broadbent, D.M., Harding, S.P., Zheng, Y.: Convolutional neural networks for diabetic retinopathy. Procedia Comput. Sci. **90**, 200–205 (2016). (In: Proceedings of Medical Image Understanding and Analysis (MIUA 2016))
38. Samet, H.: The quadtree and related hierarchical data structures. ACM Comput. Surv. **16**(2), 187–260 (1984)
39. Sutton, P.: Modeling population density with night-time satellite imagery and GIS. Comput. Environ. Urban Syst. **21**, 227–244 (1997)
40. Tadmor, E., Nezzar, S., Vese, L.: Multiscale hierarchical decomposition of images with applications to deblurring, denoising and segmentation. Commun. Math. Sci. **6**(2), 281–307 (2008)
41. Udomchaiporn, A., Coenen, F., García-Fiñana, M., Sluming, V.: 3-D volume of interest based image classification. In: Booth, R., Zhang, M.-L. (eds.) PRICAI 2016. LNCS (LNAI), vol. 9810, pp. 543–555. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42911-3_45
42. Wu, S.S., Qiu, X., Wang, L.: Population estimation methods in GIS and remote sensing: a review. J. GISci. Remote Sens. **42**(1), 80–96 (2005)
43. Zhang, Y., Zhang, B., Coenen, F., Lu, W.: Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles. Mach. Vis. Appl. **24**, 1405–1420 (2013)
44. Zhang, Y., Zhang, B., Coenen, F., Xiao, J., Lu, W.: One-class kernel subspace ensemble for medical image classification. EURASIP J. Adv. Sig. Process. **17**, 1–13 (2014)
45. Zheng, Y., Hijazi, M.H.A., Coenen, F.: Automated "Disease/No Disease" grading of age-related macular degeneration by an image mining approach. Investig. Ophthalmol. Vis. Sci. **53**(13), 8310–8318 (2012)

Exercises in Unstyling Texts: Formalisation and Visualisation of a Narrative's [Space, Time, Actors, Motion] Components

Jean-Yves Blaise and Iwona Dudek(✉)

UMR CNRS/MCC 3495 MAP, Marseille, France
{jean-yves.blaise, iwona.dudek}@map.cnrs.fr

Abstract. The research presented in this paper basis on the premise that segmenting textual content into successive situations according to four components - space, time, actors and motion – can help depicting a storyline in a way that facilitates comparative analyses across texts, and ultimately fostering knowledge discovery. The paper presents the original aim of the project and sums up the knowledge modelling choices made in order to formalise the segmentation procedure through which sequences of situations are extracted. We then present several proof of concept visualisations that facilitate visual reasoning on the structure, rhythm, patterns and variations of heterogeneous texts, and summarise how the space, time, actors and motion components are organised inside a given narrative. The approach was tested across various types of text, in three languages, and the paper details some of the potential benefits of the resulting visualisations on the specific case of R. Queneau's *Exercises in style*. The paper is concluded with a straight to the point analysis of the approach's actual weaknesses and limitations.

1 Introduction

A broad picture of the evolution of information sciences over the past decade shows that big data, meaning here big volumes of data, dynamically changing data, as well as high variety, highly heterogeneous data, has paved its way to the top of the research agenda. In parallel, availability of large collections of non-structured textual content, typically found in digital libraries, has fostered the emergence of research works clearly intermingling knowledge discovery issues with visualization issues.

A typical example of such ongoing approaches is the *ReNom* virtual library project [1]: "*…the resource associates key texts by the two Renaissance authors* [Rabelais and Ronsard] *with a system of georeferencing and a series of fact sheets corresponding to the places and characters mentioned in the texts…*". In short, corpora of texts are explored in search of Named Entities (names of places, or of people) that can then be used as a means to navigate through the texts and to "*… encourage a form of literary tourism at once fun and instructive…*".

This example does show that parameters space, time and actors can be a relevant entry point, a relevant filter, into text corpora. But the fact the approach bases on the extraction of *Named Entities* is a strong limitation: what happens when the space parameter cannot be associated with a given name, or is associated with a metaphoric name ("*the silver ribbon in the distance*" *vs.* "*on the banks of River Loire*")? What

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 28–53, 2019.
https://doi.org/10.1007/978-3-319-99701-8_2

happens when an actor remains unnamed? With what geographical map can fictional places (e.g. the city of Barchester in A. Trollope "*Barchester towers*") be associated?

Furthermore such approaches are primarily designed as means to navigate *inside* text corpora, not across text corpora, and to localise elements in the flow of the text rather than to actually segment it. They provide no systematic, abstract overview of a text's structure that would enable comparative analyses – only the number of occurrences of a given *Named Entity* in a given text can be considered as a somewhat "generic" feature.

Finally, *Named Entities* are in the above example basically associated on one hand to positions in texts, and on the other hand to a "visual" component that can be either a map (localising a name of place) or an image (portrait of a person). The visualization components are designed as end-user services, displaying univariate data, and therefore do not play a significant role in terms of analytical tasks.

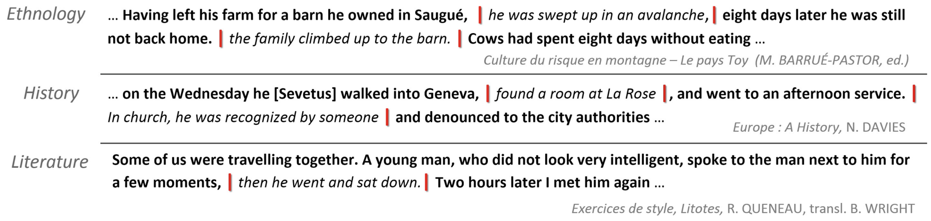
Our research is an attempt at going one step beyond in terms of genericity (encompassing texts that refer or not to Named Entities, introducing a segmentation method that covers various spatial scales, temporal markers or type of actors) and in terms of abstraction at the visualisation step (introducing a set of visual disposals allowing the analysis and cross-examination of multivariate data).

The title we have given to this contribution is a clear reference to R. Queneau's *Exercises in style*, one of the case studies that will be discussed. In that book the author demonstrates how for one same "story" (meaning one same succession of events, places, and people) there can be many ways to report how the story unfolds, there can be many ways to word out the story. But if that is so, couldn't there be a way to come back to a sort-of "root structure" of stories, enabling us to compare visually what they do have in common, and in what they differ? Wouldn't it be possible (and useful in terms of analytical tasks) to uncover and visualise that "root structure" – a storyline acting as the background of the story itself.

Our research can be seen as exploring *one* possible answer to these questions – a generic answer, applicable across various textual content, providing relatively satisfactory results, but definitely not *the* ultimate and unique answer.

It builds on the premise that a narrative can be segmented into successive or parallel ***situations*** differentiated from one another other basing on changes in ***time, space, actors,*** or ***motion***. Each situation is therefore associated with four descriptors, and a change of value of one of these four descriptors is enough to trigger a new situation.

Such *situations* act as a semantic filter, helping to analyse and compare heterogeneous texts and collections of texts basing on common metrics (Fig. 1).



Ethnology

… Having left his farm for a barn he owned in Saugué, | *he was swept up in an avalanche,* | **eight days later he was still not back home.** | *the family climbed up to the barn.* | **Cows had spent eight days without eating …**

Culture du risque en montagne – Le pays Toy (M. BARRUÉ-PASTOR, ed.)

History

… on the Wednesday he [Sevetus] walked into Geneva, | *found a room at La Rose* | **, and went to an afternoon service.** | *In church, he was recognized by someone* | **and denounced to the city authorities …**

Europe : A History, N. DAVIES

Literature

Some of us were travelling together. A young man, who did not look very intelligent, spoke to the man next to him for a few moments, | *then he went and sat down.* | **Two hours later I met him again …**

Exercices de style, Litotes, R. QUENEAU, transl. B. WRIGHT

Fig. 1. Identification of situations in heterogeneous texts [2].

Visualisations depicting sequences, rhythms, alternations of situations can then help experts and end users perform reasoning tasks on the narrative structure of texts, ranging from stylistic profiling (differences and similarities inside and across writing genres, or inside an author's works) to comparative analysis (different recounts of the same story for instance) (Fig. 2).

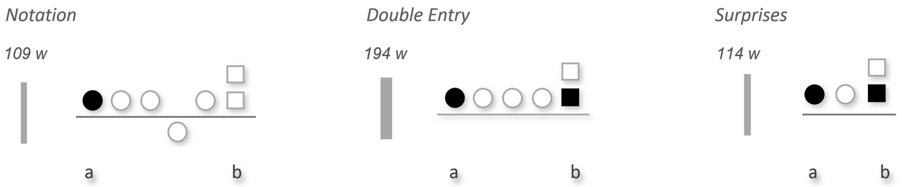


Fig. 2. A comparison of how ***situations*** unfold in time and space in three recounts of the same story (from R. Queneau's *Exercises in Style*). Note for instance that only situations (a, b) remain systematically the same across in these three versions.

The research unfolds in two sub-challenges a knowledge modelling challenge (How can we spot changes in space? What exactly makes a space to be differentiated form another – a name, a size? Who are actors - human beings only? …) and a visualisation challenge (What visual solutions could help underlining expected or unexpected patterns inside or across texts?).

The paper is structured as follows: Sect. 2 introduces the reason to be of this research - it discusses the notion of *situation* that is at the heart of the approach. In Sect. 3 we position our contribution with regards to existing approaches in the fields of visual analytics on one hand, and of text analysis on the other hand. Section 4 details our choices in terms of knowledge modelling, *i.e.* how the *space/time/actors/motion* components are used in the segmentation of textual content. Section 5 then presents a series of experimental visualisations corresponding to alternative combinations of variables and ultimately to alternative reasoning tasks. We then illustrate the potential benefits of these visualisations in terms of knowledge discovery for text analysis on one specific case: R. Queneau's *Exercises in style*. Section 6 describes the implementation and evaluation efforts carried out up to now. In Sect. 7 we pinpoint strengths and weaknesses of the approach, and in particular challenges ahead if wanting to apply the approach on a large scale. Finally, a short conclusion section sums up what we think can be considered as fruitful feedbacks from this study.

The paper is an extension of a preliminary contribution entitled *StorylineViz: A [Space, Time, Actors, Motion] - Segmentation Method for Visual Text Exploration* [2].

2 Research Issue

2.1 Origin of the Research

There is naturally a large range of features researchers may want to extract from text corpora, and analyse through visual means. Some are clearly *structure-related*, like in Marshman's [3] comparative analysis of lexical knowledge patterns. Others by contrast focus on spotting *topics* like Sabol's topical-temporal maps [4], a visual metaphor allowing an interactive analysis of how prominent topics in large collection of news releases change over time.

So why did we choose to focus on extracting the ***spatio-temporal content*** of textual data? The idea came as a natural continuation of years of research conducted on the architectural and urban heritage. Our usual concern, intersecting InfoVis (Information Visualisation) and Heritage sciences is analysing and visualizing architectural transformations, from the point of view of morphology (changes of shapes), from the point of view of chronology (duration, dating of changes), and from the point of view of events and people (correlation of data about changes) [5, 6]. In short, we pull together a large amount of heterogeneous historical evidence, implement ad hoc information systems and ultimately provide researchers with means, inspired by the InfoVis legacy, to analyse this evidence visually.

The input we handle is therefore historical evidence: *hints* about space (e.g. an edifice, a garden, a marketplace), time (e.g. a period of construction), events (e.g. a war, a plague) and people (e.g. a ruler, a builder, an owner). At the root of our research processes there are physical (remains), visual (iconography) or ***written*** testimonies and pieces of data, that we consider from the point of view of what they tell us of a space, of a moment in time, of events and their impacts, of people that act in that space at that time.

A significant part of the historical evidence we use is extracted from texts, ranging from inventories to travel diaries and historical research, used for instance to "anchor" events and actors in time, space and context:

> *"… On Saturday, 12 August 1553, a fugitive from the holy inquisition rode into the village of Louyset…"* (N. Davies, *Europe: A History*, Pimlico, London 1997, p. 493)

At the end of the day hints are recorded as corresponding to a given place, a given time, a given set of actors that altogether form a specific *state*.

Recounting and analysing an evolution can then be done by comparing successive "states", opening on a very general notion of '*path*', understood as a series of situations leading from an initial state to a final state. This series is consistent or not in terms of spatial scale or quality of the information describing situations. It can be continuous or not (*i.e.* including or not temporal breaks).

2.2 From Reasoning on *States* to Reasoning on *Situations*

The above notion of path can be used to interpret and structure (i.e. segment according to consistent division lines) a variety of heterogeneous historical evidence: travel diaries, witness reports, inventories, iconographic material, *etc*. But could it act as a potential semantic filter far beyond its initial field of concern - historical evidence?

When going through historical evidence quite often both place and time are likely to be partially, when not poorly, described - a document will for instance mention something occurring "*on street* ***A*** *at the beginning of spring*". Neither space, nor time are consistently defined inside sources, and across sources (varying precision, varying granularity). In that context a 'path' - understood as a series of potentially ill-defined situations - is obviously closer to the content of a *narrative* that to highly structured data sets handled in route calculations offered by GPS applications for instance. Hence the attempt we present in this paper to try and see to which extent such an approach to text segmentation could be fruitful, beyond its initial context of emergence.

2.3 Segmenting Narratives as Series of Situations

StoryLineViz should be understood as a proof of concept study that aims at developing a generic approach to narrative analysis, supporting the identification and visualisation of significant patterns inside textual data, and ultimately knowledge discovery and sense-making. Narratives as seen from that general point of view are strongly heterogeneous (from whole texts to just series of facts, from a book or collection to a few paragraphs). In addition, they can be contradictory or conflicting (different recounts of a series of events) or transformed (typically by translations). As of today they are often categorised (a play, a travel diary, an eye-witness report) and analysed from an expert's point of view (linguistics, literature, history, *etc.*) but hard to synthesize and to compare to one another.

In this contribution we propose an approach in which a narrative is segmented in a series of situations in ***ordinal time*** (*i.e.* only the order of appearance of situations is defined: situation ***A*** occurs before situation ***B***, but neither ***A*** nor ***B*** need to be actually dated). A situation is differentiated from another basing on the variation of one of the four following parameters: time, space, actors, and motion.

Our approach's core objective is to facilitate visual reasoning on the structure, rhythm, patterns and variations of texts in order to enable comparative analysis and to summarise in a clear-cut manner how the space/time/actors/motion components are organised inside and across narratives.

Quantitative and qualitative parameters can then be taken into account, allowing the association of causal or contextual indicators. The segmentation procedure is seen as a common ground between varieties of narratives. It aims at facilitating visual reasoning on the structure, rhythms, patterns and variations inside narratives or across collections of narratives. If proven workable the approach opens a number of application scenarios, among which:

- comparing oral or written recounts of the same series of events by different witnesses,
- comparing different recounts of the same itinerary (e.g. what differences can be spotted in the way pilgrims walk the Camino de Santiago today, and before, back to the initial *Codex Calixtinus*),
- supporting the identification of trends, patterns, evolution in writing genres in an edutainment-like approach (e.g. To which extent do classic Greek theatre plays and their 20th century reinterpretations overlap with the famous "unity of space, time and action" rule?),
- uncovering differences in the interpretation of texts by different readers.

At the end of the day, the approach can also be seen as an attempt to step out of discipline-specific frameworks so as to promote sort of “universal”, comparison-enhancing, metrics of narratives. However it should be said right away that this research makes no claim in the context of natural language processing or massive content analysis – what is presented in our contribution is basically an attempt to try and see if the specific segmentation bias we introduce could stimulate further research and lead to unthought-of observations on the structure of narratives.

2.4 The Concept of Situation: Legacy, and Open Challenge

The idea that a narrative is (at least in part) composed of successive *situations*, corresponding to *space/time/actors/motion* components, is definitely not new.

J.R.R. Tolkien’s *The Hobbit* starts as follows: “… *In a hole in the ground there lived a hobbit.*” The narrative is triggered in time, indications about space (a hole), actors (a hobbit) and motion (no indication of movement in that sentence) are given.

The idea that a narrative can be presented, ***including visually***, as a series of situations is also far from being new. Italo Calvino in his remarkable *Collection of Sand* comments on the “figurative narrativity” of Trajan’s Column, along which Trajan’s two wars in Dacia are recounted situation by situation through engravings. Situations follow one another (ordinal time) along a 200 m long spiral, going from bottom to top. Each bas-relief corresponding to one situation contains indications about space and actors - Trajan before the imperial tent, a legionnaire digging a ditch, and so on.). In another essay Calvino comments on the development and use of *ribbon maps* over time - a map including time and space, a map representing a path. A beautiful example of such maps is the mediaeval (12th century) ribbon map representing the “London To Jerusalem” itinerary by Matthew Paris, on which each stopover is represented by an outline of the city [7].

The idea to depict situations in a less figurative manner than in the above examples ins not new neither. For instance *Historical centographs* developed during the 19th century as a mnemonic system [8] introduce a level of abstraction, and an ambition for visual comparative analysis, that are much closer to our approach. *Centographs* represent time (ordered time model) though grids of squares: a 10 × 10 grid represents 100 years. Each square (representing one year) is then subdivided in a 3 × 3 grid, with each of the nine sub-squares representing a “variable” that can be chosen freely (lifetime of a person, event such as a war going on, *etc*.). Users of the system would fill in blank templates in order to underline co-occurrences in an attempts at facilitating memorisation tasks for students. This example does not precisely correspond to the notion of situation in narrative, but the visual solution can act as food for thinking: *space/time/actors/motion* are four “variables” that we observe in order to trigger a “situation change”.

More recent, and situated in terms of graphic language somewhere in between the figurative nature of *ribbon maps*, and the strict abstract nature of *centographs*, is the *Home to School* diagram by Yabuuchi [9] where an itinerary is depicted as a series of situations each of which being associated with a set of “qualitative” variables (sound, type of landscape, type of vehicles, *etc*.). The resulting visualisation is a very elegant linear diagram looking more or less like a musical stave with each line bearing a variable, and the overall diagram clearly underlining consistent sequences in the overall itinerary, major trends, and exceptional situations.

What can be said is that these examples do back up the idea that indeed a narrative can be segmented into a series of situations, reduced to sequences that somehow summarise the story, provide outlines, stop-points. But they give no hint at all on how to segment narratives, on which division lines should be used to decide whether or not a new situation has to be reported. Furthermore, they give no hint on how features corresponding to each situation (e.g. space, time, actors, and motion) can be compared to one another. For instance in the above example of Trajan's Column the initial situation, from the point of view of space is "*the landscape of a fortified Roman Town*", and the next situation "*Roman soldiers crossing the Danube on a pontoon bridge*". How do these "spaces" relate to one another? In Sect. 4 we propose a strategy for associating each feature with a formal grid of descriptors designed as a mean to allow for cross-examinations: the "*fortified Roman Town*" spatial indication would be associated with scale ***7***, and the "*Roman soldiers crossing the Danube*" with scale ***8***.

In other words our contribution is definitely not in pushing forward the concept of *situation*, but in the interplay between formal segmentation rules (enabling the identification of situations) and interactive visualisations (enabling a user-side analysis of situations and sequences of situations). Ultimately the challenge addressed here is to test a segmentation into *situations* that potentially says something about the structure of the narrative itself, or about the producer of the narrative.

3 Scientific Context

Open access to massive textual content, typically as found in digital libraries, has fostered the emergence of research works intermingling knowledge management, visualization, and language processing issues. In this contribution we focus on large non-structured texts. Unlike when handling structured data sets, working on large texts, today often made available in large open access repositories such as *Gallica*, introduces specific challenges. Oelke [10] summarizes some of them: quantity (amount of words), polysemy (of words, references, literary imagery), flexibility (of rules in natural languages), interpretation (use of a predefined knowledge of the world by humans).

A typical example where space and actors are extracted from narrative texts is the CHAPLIN (CHAracters and PLaces Interaction Network) tool [11] - following a user-monitored extraction of terms graphs are produced that represent connections between places and people. Another significant example, this time focusing on temporal aspects, can be found in [12] - visual analyses of sentiments and character interaction in the flow of a fiction – an approach closer to what will be discussed in this paper but basing here again on the appearances and co-occurrences of named characters in chapters.

Said briefly, there is a move towards bridging the gap between on one hand linguistics-based approaches – i.e. for instance spotting markers of cause-effect relations in text corpora, as in [13] – and on the other hand information visualisation approaches – i.e. for instance *tileBars* for document visualisation [14], or basic *wordclouds*. Hence supporting text analysis through visual means has become a hot research topic in the field of visual analytics (VA), a field described in its early days by

Thomas and Cook as "*focusing on analytical reasoning facilitated by interactive visual interfaces*" [15].

Our study proposes an approach that centres on semantic aspects, applicable across collections of texts. It builds on the idea that visualisation can help users explore, analyse and cross-examine textual documents. This idea is backed by research works covering a wide range of issues: *VisRA* tool [16] focuses on readability analysis, *VarifocalReader* [17] focuses on multi-layer visualisation/navigation and interactive annotation, *POSvis* [18] on relationships and co-occurrences in the flow of a text, Wanner's approach [19] digs in the notion of opinion and sentiment in book rating.

Those examples share a common mantra: *human analysis of textual content and sensemaking in large and/or complex textual data sets can be facilitated by adapted abstract visualisations*. They also share a common statement: *full automatic algorithms can hit their limit when facing complex texts*.

Accordingly, our study does relate to the above research works in terms of scientific context, but it clearly leaves aside the NLP (Natural Language Processing) issues. We shall in this contribution focus on the knowledge modelling step on one hand, and on the visualisation step on the other hand. Mainstream research works at the intersection of VA and NLP have been investigating approaches that strongly rely on a line per line, word per word analysis of textual content: statistical approaches (e.g. occurrences of words, lengths, types of words), Named Entities Recognition (NER) related approaches (e.g. user selections of words, ontologies, opinion indicators), machine-learning approaches (e.g. extraction of significant linguistic patterns). In all these cases, language itself - *i.e.* the occurrences, positions, lengths, relations of words and sentences - is at the heart of a discipline specific analysis.

By contrast our approach builds on a segmentation bias that is:

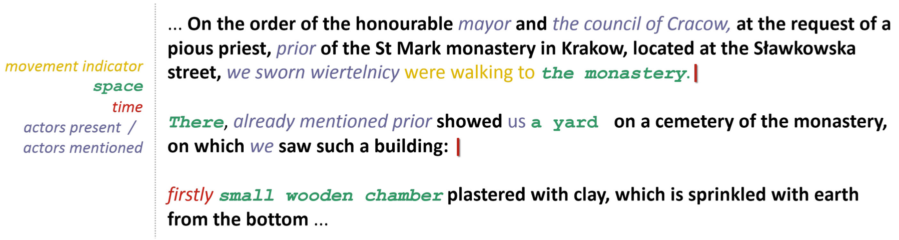
- neutral - allowing for a discipline-independent cross-examination of texts,
- unrelated to text features such as lengths (a new situation can occur inside one sentence, or after three pages),
- focusing on supporting visual comparisons of rhythms and sequences, at user-chosen aggregation levels.

As will be discussed in Sect. 7, we do acknowledge that the language processing step remains at this stage of our research an unaddressed issue. The segmentation of texts used in the study has been done manually: it could be seen as a weakness in terms of significance and reusability of the approach.

We believe that before any attempt at "automatizing" language processing it is key to formalise a robust, insight-gaining, unambiguous segmentation protocol, and to evaluate in what the visualisations can be beneficial. Accordingly we consider that our study can contribute to pinpointing a new research path, at a time when the focus is often put on the processing of massive data sets.

4 Method

We introduce a text analysis method that builds on the identification of quadruplets of components: actors, space, time and motion. These components are used to segment a narrative and translate it into sequences of situations in ordinal time (only the order of events considered) (Fig. 3).



movement indicator
space
time
actors present /
actors mentioned

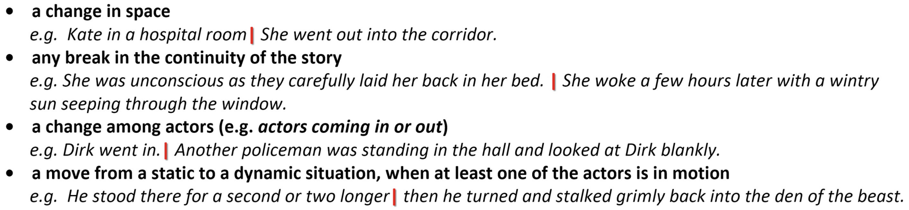
… On the order of the honourable *mayor* and *the council of Cracow,* at the request of a pious priest, *prior* of the St Mark monastery in Krakow, located at the Sławkowska street, *we sworn wiertelnicy* were walking to the monastery. |

There, *already mentioned prior* showed us a yard on a cemetery of the monastery, on which *we* saw such a building: |

firstly small wooden chamber plastered with clay, which is sprinkled with earth from the bottom …

Fig. 3. Segmentation into situations - four indicators (*Actum feria sexta ante Fabiani et Sebastiani [191] anno Domini 15*96) [2].

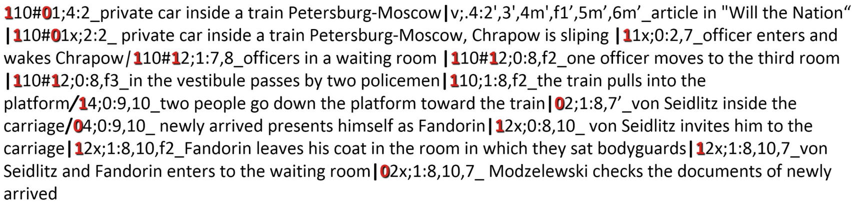
A situation is basically a sort-of token, resulting from the segmentation procedure. However we are here far from a segmentation at the word or sentence level: situations are determined by changes of values in a quadruplet of descriptors (space, time, actors, and motion). A change of one of the four descriptors introduces a new situation - (Fig. 4). Situations occurring in the past of the story (e.g. reminiscences - narrating past experiences) are differentiated from those occurring in the course of the story.



- **a change in space**
  e.g. Kate in a hospital room | She went out into the corridor.
- **any break in the continuity of the story**
  e.g. She was unconscious as they carefully laid her back in her bed. | She woke a few hours later with a wintry sun seeping through the window.
- **a change among actors (e.g. *actors coming in or out*)**
  e.g. Dirk went in. | Another policeman was standing in the hall and looked at Dirk blankly.
- **a move from a static to a dynamic situation, when at least one of the actors is in motion**
  e.g. He stood there for a second or two longer | then he turned and stalked grimly back into the den of the beast.

Fig. 4. A segmentation procedure ending in the identification of independent situations basing on changes in space, time, actors or motion: example of application to D. Adams' *The Long Dark Tea-Time of the Soul* [2].

Situations are identified at this stage through a manual annotation and segmentation process - a dozen of texts ranging from literature to ethnology have been tested, covering three languages. Each situation is associated manually with a value for each of the four descriptors, and with a short paraphrase summarising "what happens". The four values are translated into an alphanumeric code comprising indicators for each of four parameters and separators that allow for a processing of the information (Fig. 5).



```
110#01;4:2_private car inside a train Petersburg-Moscow|v;.4:2',3',4m',f1’,5m’,6m’_article in "Will the Nation“
|110#01x;2:2_ private car inside a train Petersburg-Moscow, Chrapow is sliping |11x;0:2,7_officer enters and
wakes Chrapow/110#12;1:7,8_officers in a waiting room |110#12;0:8,f2_one officer moves to the third room
|110#12;0:8,f3_in the vestibule passes by two policemen|110;1:8,f2_the train pulls into the
platform/14;0:9,10_two people go down the platform toward the train|02;1:8,7’_von Seidlitz inside the
carriage/04;0:9,10_ newly arrived presents himself as Fandorin|12x;0:8,10_ von Seidlitz invites him to the
carriage|12x;1:8,10,f2_Fandorin leaves his coat in the room in which they sat bodyguards|12x;1:8,10,7_von
Seidlitz and Fandorin enters to the waiting room|02x;1:8,10,7_ Modzelewski checks the documents of newly
arrived
```

Fig. 5. Example of the alphanumeric code resulting from the annotation phase (in red, the code corresponding to the motion indicator - 0 static 1 dynamic – example from *The Death of Achilles* by B. Akunin) [2]. (Color figure online)

Situations can also be grouped by predefined sequences such as chapters (or any other main division of a document). The way each component is defined and structured is detailed in the following subsections. At the end of the analysis phase the text under scrutiny is entirely transformed into sequences of situations as they occur in the narrative. Sequences are then translated into a visual language.

4.1 Space Parameter

The space parameter defines where the action takes place (*i.e. Where does the action begin? Does it continue in the same location? Are the subsequently cited places well identified or in vaguely mentioned locations? Are their many quick changes of space? Are these changes related with a jump in time, a flashback for instance?*). But ‘space’ as geographers, historians, architects, or ethnologists picture it is far from being one and only one notion. It can be described quantitatively (positions, size, exact morphology) or qualitatively (through linguistic indicators, or a relation to a Named Entity, for instance a toponymy like in [20]).

In the context of this study we need to spot in the flow of a narrative the moments when a change of space occurs, and therefore leads to a new situation (whether spaces are associated with a given named entity - *e.g. Paris*, or are present in the flow of part-of-speech – *e.g. in the second cellar*). Detecting such changes implies defining unambiguous lines of division between spaces.

To do so, we reinterpret the concept of scale (in accordance with previous research on spatio-temporal information retrieval [5, 21, 22]. What is meant by scale is not a map’s numerical ratio, but the idea that spaces can be classified according to alternative spatial granularities.

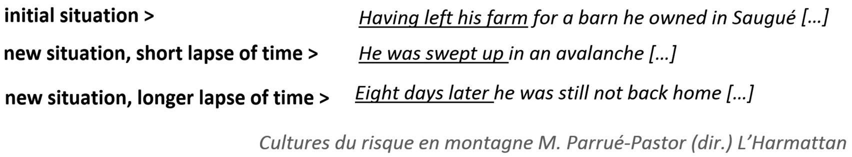
Our model of space includes 16 indicators (3 non-spatial descriptors and 13 scale identifiers). The non-spatial descriptors concern the situations where space is not clearly assessed (metaphorical descriptions, undefined space, space is not present) –in other words non-spatial descriptors help dealing with incomplete, ill-defined, or simply missing spatial information.

The thirteen scale identifiers are organised into six groups (e.g. in and around a building, public spaces, open land). An additional parameter is taken into consideration: primary vs. nested spaces. ***Primary spaces*** correspond to ‘simple’ situations (e.g. Jane is in her room, Jane is walking in the garden). ***Nested situations*** appear when actors are inside vehicles or objects that can move or be moved inside a primary space (e.g. Jane is travelling by train.).

4.2 Time Parameter

The time parameter corresponds to the when question: it explains the story's development over time (*e.g.* continuous progression from present to future, regressive present-to-past development, multiple changes of time, *etc*.).

The time model builds on the notion of ordinal time [23]: situations are analysed from the point of view of an order of appearance (before/after) in the flow of the narration, but neither quantified nor anchored (Fig. 6).



initial situation > *Having left his farm for a barn he owned in Saugué […]*

new situation, short lapse of time > *He was swept up in an avalanche […]*

new situation, longer lapse of time > *Eight days later he was still not back home […]*

Cultures du risque en montagne M. Parrué-Pastor (dir.) L'Harmattan

Fig. 6. Change of situations - temporal disruptions.

A qualitative assessment of time continuity is associated to each situation change (lapse of time separating a situation from the next one).

Successive situations are identified in the order of the narration (as the story unfolds) as belonging to the ***present*** of the story or its ***past*** (things having occurred "before the present of the story"). Situations can also be tagged as being ***parallel*** (occurring at the 'same' time) (Fig. 7).

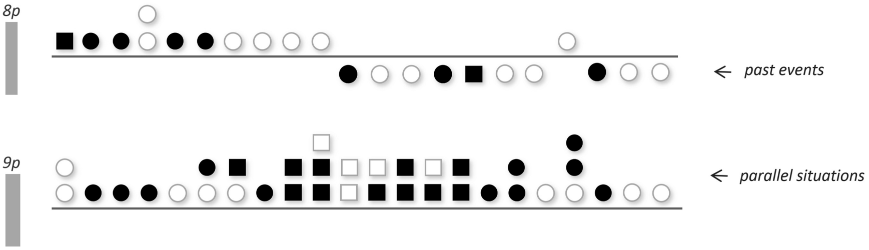


Fig. 7. Top, ***past situations*** are represented below the horizontal line. Bottom, ***parallel situations*** are represented by graphic elements "piled" one over the other above the horizontal line [2].

Additional indicators are used to further describe parallel situations (actors mutually aware of one another or not, typically), or to identify customary behaviours (occurring repeatedly).

In the context of this study we need to spot in the flow of a narrative the moments when a change of space occurs, and therefore leads to a new situation (whether spaces are associated with a given named entity - *e.g. Paris*, or are present in the flow of part-of-speech – *e.g. in the second cellar*). Detecting such changes implies defining unambiguous lines of division between spaces.

4.3 Actors

Actors are yet another trigger of situation change. They may be individuals, well defined groups of people, but can they also be indistinctly specified groups (*e.g.* a crowd), things (*e.g.* thinking machines), or animals? We here need to disambiguate the very concept of actor: are ants mentioned in B. Werber's *Empire of the ants* actors?

Our strategy is to consider actors as a being or a consistent group of beings, real or imaginary creatures or entities, fitted with the ability to make choices and to act. Actors may be human beings, but also gods (*e.g.* Zeus, Dionysus), thinking machines, androids, animals (*e.g.* the wolf in Little Red Riding Hood), and so on. The description of actors is then fine-tuned. Actors physically engaged in a situation (*i.e.* present) are distinguished from actors that are only mentioned (*e.g.* in a conversation, or in thoughts), individual actors are distinguished from consistent groups either identified (*e.g.* the Celts) or not (*e.g.* a crowd) (Fig. 8).

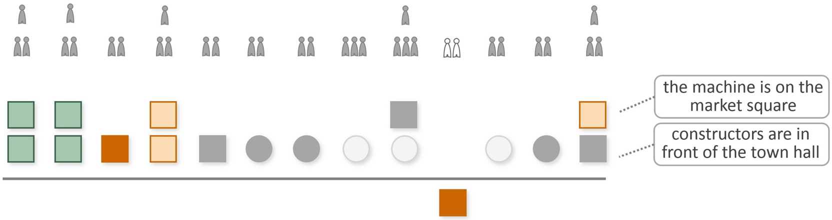

the machine is on the market square
constructors are in front of the town hall

Fig. 8. Actors appearing in each situation of S. Lem's *Trurl's Machine*. Situations are read from left to right. Here Trurl and Klapaucius, the two engineers (bottom part of the lines showing actors as silhouettes) are being chased by Trurl's machine gone mad (top part of these lines, one silhouette alone). A reference to past events is made (orange square below the horizontal grey line), and that past situation concerns two actors not present but mentioned (white silhouettes). (Color figure online)

Finally, major events concerning actors can also open up on a situation change – a severe injury, or a death of an actor needs to be reported.

4.4 Motion

Finally, motion is also a key element in the definition of a situation (only the motion of actors is considered). Motion is important to state since it helps unveiling spatial and temporal continuities or discontinuities in the narrative. An intensive use of motion indicators in a text may characterise writing genres (*e.g.* logbooks), may underline recurrent stylistic elements (*e.g.* a speed chase with the police), stylistic characteristics of an author, differentiate acts inside one play, help understanding changes in space, and so on (Fig. 9).

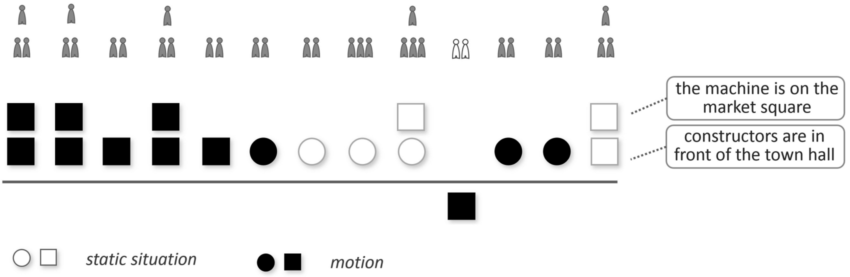

the machine is on the market square
constructors are in front of the town hall
static situation
motion

Fig. 9. A partial view of motion analysis visualisation corresponding to S. Lem's *Trurl's Machine*. Light grey elements indicate static situations (e.g. the engineers discuss with the mayor of a town in which they sought refuge).

Naturally we need to be clear on what we mean by motion. The strategy is to focus on movements that introduce a change of location but not necessarily a change of space (*e.g.* someone is walking down a street). From the point of view of this criterion situations may then be classified as ***static*** or ***dynamic***. A ***dynamic situation*** implies the motion of at least one of the actors, motion understood as moving in space (*e.g.* walking, marching, strolling, running, driving a car…).

5 Visual Solutions

Our approach bases on the idea that interactive visual interfaces can help various target users perform reasoning tasks in application fields ranging from expert analysis to education or cultural mediation. Accordingly visualisation is a key component of the study, both in the understanding of a given narrative's "spatio-temporal profile" and in fostering comparisons inside collections of texts. What is meant here by the term visualisation is made clear by Sabol [24]: (a) graphical representation of data, information and knowledge visualisation, (b) using the human visual system, supported by computer graphics, to analyse and interpret large amounts of data, (c) visual representation to aid cognition.

Depending on the parameters a user may choose to privilege (*i.e.* space, time, actors or motion), different visualisations are proposed. We detail them in the following subsections. All of these visualisations share some common design principles and a graphic language that we try to adapt to a human's visual apparatus limited number of pre-attentive features [24]:

- Situations are represented one by one and aligned as they occur in the original text (left to right, or top-down),
- Each situation is represented by an interactive symbol (a multidimensional icon). Shape, colour and position are used to transfer visually the information on each situation,
- A rephrasing of the actual text corresponding to each situation is available on user demand,

- Parallel situations, *i.e.* situations co-occurring in time, are grouped and represented together,
- Actors are visualised on user demand, with colours differentiating the nature or type of actors (actors present, mentioned, injured, or groups of actors).

Situations are grouped by sequences (chapters or other grouping mechanism adequate for a particular writing genre) in order to grab more easily an understanding of the text's structural features.

5.1 Spatial Sequences Visualisation

In the ***spatial sequences*** visualisation situations are represented in ordinal time from left to right along horizontal bars. Each horizontal bar corresponds to a sequence of situations. All reminiscences are situated below horizontal bars (Fig. 10 $\boldsymbol{b_1}$, $\boldsymbol{b_2}$, $\boldsymbol{b_3}$). Colour and shapes are used to differentiate the occurrences of various spatial scales.

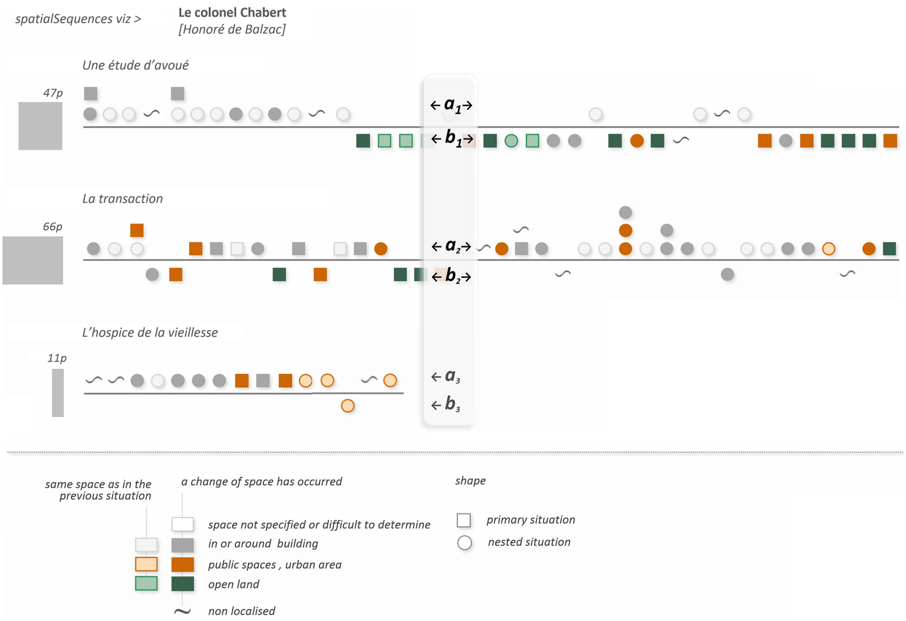


Fig. 10. Organisation and legend of the *spatial sequences* visualisation. Bottom, legends of the visualisation - colours correspond to ranges of scales. Squares and circles differentiate nested spatial configurations (e.g. driving a car in a city) from primary configurations (e.g. walking in a city). Top, a partial view of the spatial sequences visualisation corresponding to Balzac's *Colonel Chabert*. Note for instance the contrast between spatial location of present (***a1***) and past (***b1***) of the story in chapter one (colours), or the quasi-absence of past events in chapter 3 (***b3***) [2]. (Color figure online)

5.2 Motion Analysis Visualisation

The ***motion analysis*** visualisation uses the same general organisation as the previous: situations are represented in ordinal time from left to right along a horizontal bar. But here the focus is put on the motion component of the model: colours and transparency representing different types of space are replaced by black-and-white motion indicators.

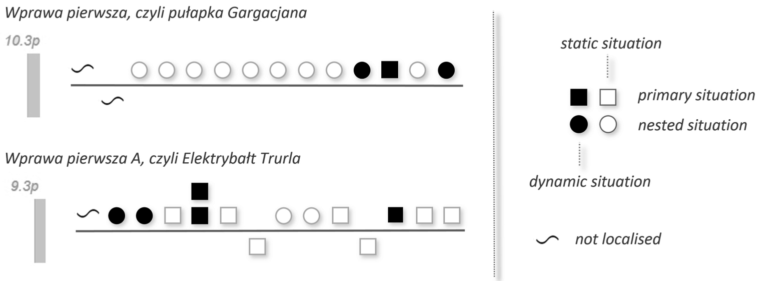


Fig. 11. Organisation and legend of the *motion analysis* visualisation. Left, a partial view of the visualisation corresponding to S. Lem's *Cyberiada*. Note for instance the long sequence of static, nested situations in chapter 1. Right, legend of the visualisation [2].

This visualisation is used to differentiate static and dynamic situations, thereby better underlining in particular rhythms inside a text (Figs. 7, 9, 11 and 14).

5.3 Temporal Continuity

The ***temporal continuity*** visualisation focuses on assessing visually to which extent the story unfolds without interruption in time (Fig. 12).

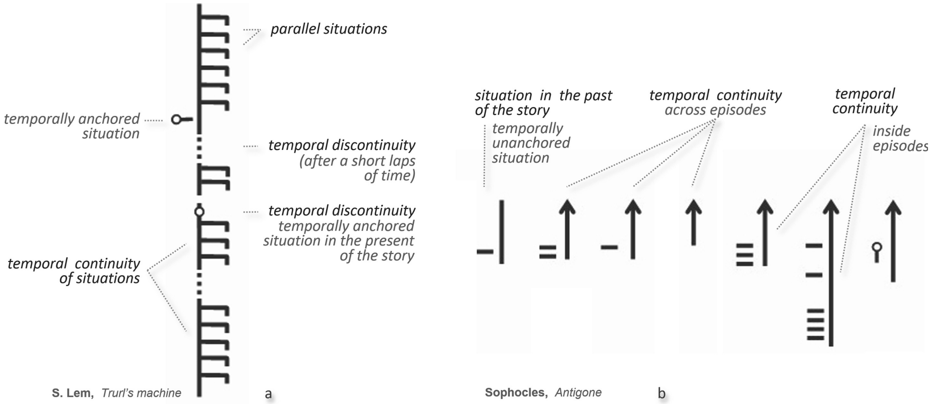

parallel situations
temporally anchored situation
temporal discontinuity (after a short laps of time)
temporal discontinuity temporally anchored situation in the present of the story
temporal continuity of situations
S. Lem, Trurl's machine
a
situation in the past of the story
temporally unanchored situation
temporal continuity across episodes
temporal continuity
inside episodes
Sophocles, Antigone
b

Fig. 12. The *temporal continuity* visualisation: (a) applied to S. Lem's Trurl's Machine. The visualisation shows an intensive use of parallel situations, and spots three lapses of time disrupting the temporal continuity, (b) applied to Sophocles' Antigone, the visualisation illustrates the unity of time pattern - a typical example of classical unity of time rule for drama.

Each sequence (*i.e.* chapter, episode, *etc.*) is here represented as a vertical line. A line topped with an arrow shows a temporal continuity with a previous situation. Small horizontal lines distributed on the left side of the vertical line correspond to situations occurring in the past of the story. Parallel situations are identified by symbols positioned on the right side of the vertical line. The vertical line is disrupted by various symbols in cases of temporal discontinuity (different symbols are used to represent short lapses of time, jumps in time, temporally unanchored events, *etc.*).

5.4 Spatio-Temporal Continuity

The ***spatio-temporal continuity*** visualisation builds on the same design as the previous, but adds symbols representing the space parameter. Whereas in the spatial sequences visualisation (Sect. 5.1) we only deliver an indication about the group of scales corresponding to a situation, we here allow for a visual coding of each of the thirteen individual scales. Fine-grain differences can be made for instance to differentiate a situation occurring inside a building from a situation occurring in a building's courtyard, or in a flat forming part of the building (Fig. 13).

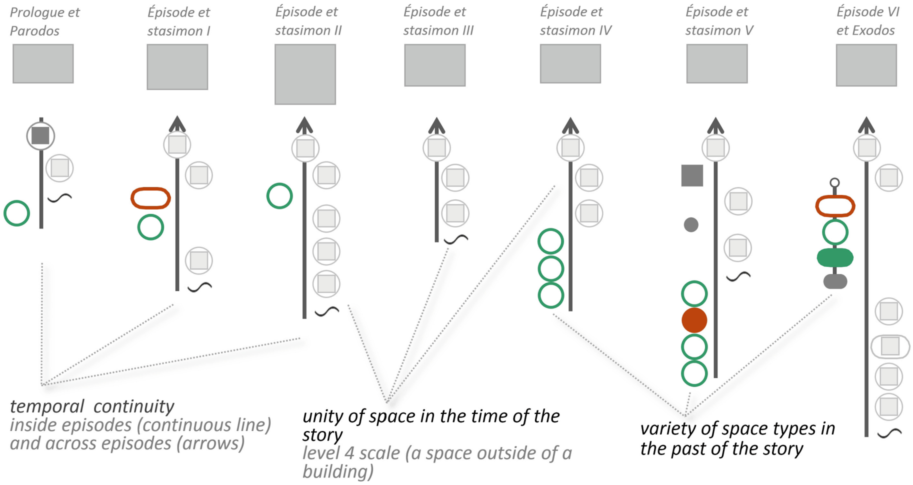

Prologue et Parodos
Épisode et stasimon I
Épisode et stasimon II
Épisode et stasimon III
Épisode et stasimon IV
Épisode et stasimon V
Épisode VI et Exodos
temporal continuity
inside episodes (continuous line)
and across episodes (arrows)
unity of space in the time of the story
level 4 scale (a space outside of a building)
variety of space types in the past of the story

Fig. 13. Spatio-temporal continuity visualisation corresponding to Antigone of Sophocles. Note contrast in terms of space between the past of the story (symbols on the left of the vertical lines) and the present of the story (symbols situated on the vertical time-line and on the right of it). Note also that in the present of the story space remains unchanged (*in front of the palace*).

5.5 Illustrating the Approach on a Case Study: Exercises in Style

Raymond Queneau's *Exercises in Style* is a 1947 book in which the author retells the same story 99 times, each time in a different "style". The word "style" should however here be interpreted as something of an over-simplification when reading the exact titles given to the versions: "metaphorically", "precision", "logical analysis", "philosophic", "official letter", *etc*. What Queneau does really is demonstrating there is a distance between the events reported and the wording. Briefly said the story is this of someone who gets on a bus, witnesses an altercation between a long-necked man and another passenger, and then notices this same long-necked man two hours later close to a railway station and getting advice on adding a button to his overcoat. Because it clearly questions the distance between facts and reports of facts, between what makes “style” a layer “above” the meaning of what is said, Queneau’s work was obviously a promising test bench for our approach.

And indeed, as is shown in the examples proposed below, the series of events reported in each of the 99 versions is not exactly the same: the storyline remains, but there are here and there situations that are added, withdrawn, or modified in terms of what is known of each situation. Because the method we explore involves exclusively relations between time, space, actors and movement, we naturally make no claim that the segmentation procedure we propose, and the visualisations we have tested, do cover the entire subject matter of Queneau’s illustration of the distance between a storyline and its wording.

But the following examples show our approach does help pointing out significant resemblances and contrasts between versions, trends and exceptions in the whole collection of 99 versions – *i.e.* encourages analytical reasoning on textual content.

Highlighting Differences. This first series of visualisations show the approach does help highlighting differences between the various versions of the story. Some sharp contrasts can be noticed in terms of quantities (number of situations, number of situations that include motion, *etc.*) but also in terms of “stylistic figures” (use of recollections, parallel situations, metaphors, *etc.*) (Fig. 15).

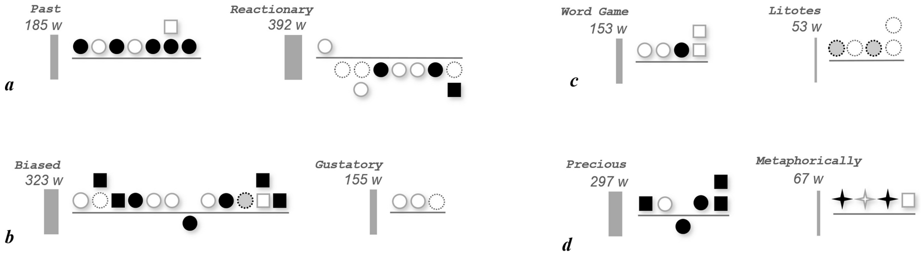


Fig. 14. *Motion analysis*. Two by two comparisons underlining differences between versions: (a) story told in the present *vs.* as a recollection; (b) high contrast in number of situations; (c) spaces named *vs.* unspecified; factual *vs.* metaphorical recount.

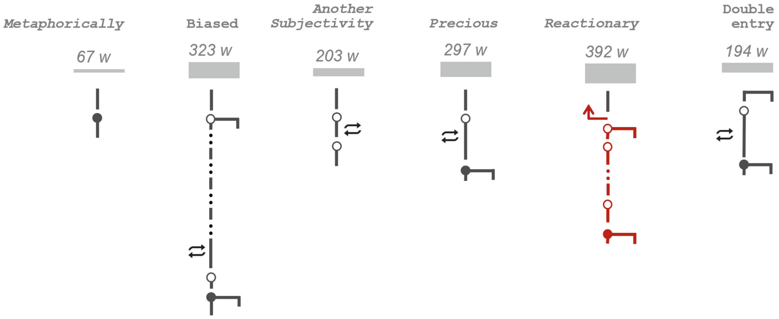


Fig. 15. A comparison of temporal patterns across 6 versions showing a strong variety.

Spotting Groups and Patterns within Groups. The fact that stories differ from one another is not surprising – the visualisations in that case basically back up something that is rather intuitive for readers, and give some "factual" basis to that intuition. But the fact that within the collection there are consistent groups of stories (in terms of space/time/actors/motion components) and that within these groups there are some specific patterns just cannot be noticed naturally, intuitively. Yet such groups, and patterns, are clearly spotted in the examples below (Figs. 16 and 17).

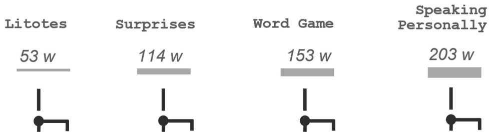


Fig. 16. Four *temporal continuity* visualisations corresponding to four versions that are strictly identical from the point of view of time, thereby forming a consistent group independently of the lengths of the stories (counted in words, represented by the greyish rectangle above the glyphs).

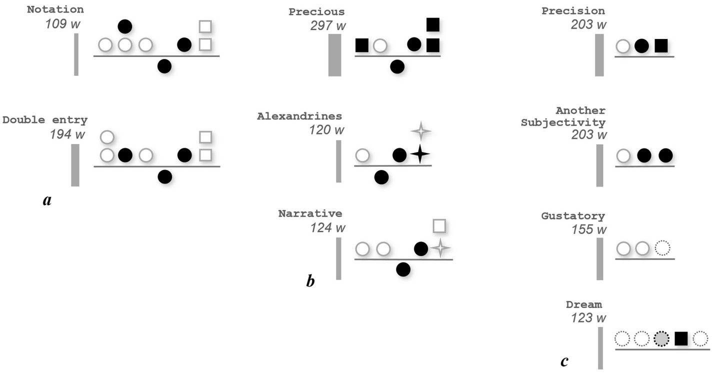


Fig. 17. *Motion analysis* visualisations corresponding to three groups: (a) two versions with almost the same number and type of situations; (b) different types of situations but a final situation always parallel, and a situation in the past always with motion; (c) a group of version with a rather limited number of situations, all in the present of the story, and no parallel situations.

Different Stories, but Invariant Features. The visualisations also help spotting across the set of stories what can be called invariant features – typically the initial and final situations emerge as such (Fig. 18).

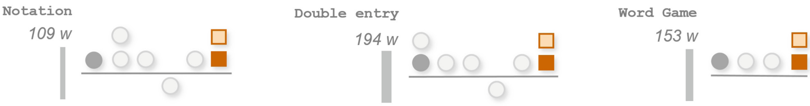


Fig. 18. These *spatial sequences* visualisations show a clear invariant feature in three versions: final parallel situations with the same indicators of space (both situations in *cour de Rome*).

Number of Words, Number of Situations. One could probably think that the more words in the story, the more chances to meet new situations. Or one could think these quantities are unrelated. The visualisations show some interesting resemblances between versions in terms of number of situations and number of words. But they also clearly show that the number of situations is not correlated to the number of words, rather to stylistic choices (presence of verbose descriptions of places or people for instance, or of feelings and thoughts) (Fig. 19).

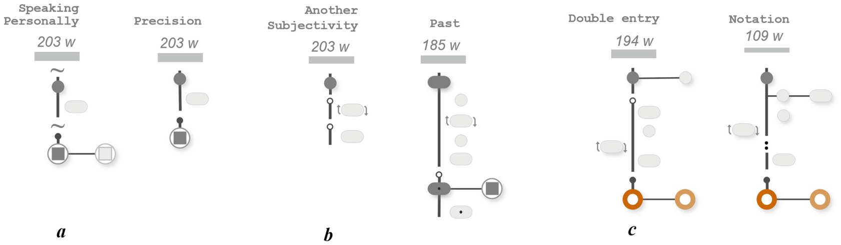


Fig. 19. Six *spatio-temporal continuity* visualisations: (a) same length, and almost same number of situations; (b) longer text on the left, higher number of situations on the right; (c) almost similar diagrams but contrasting lengths.

Fine-Grain Comparative Analyses. A number of fine-grain one to one comparisons and analyses can be carried out on versions of the story by observing for instance the spatial granularity used to tell the story, the temporal disruptions introduced, the exact moment of appearance of the same event in the various versions, *etc.* (Fig. 20).

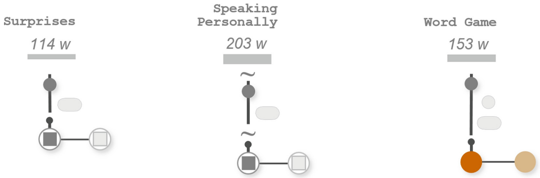


Fig. 20. *Spatio-temporal continuity* visualisations. The number and type of situations in version *Surprises* and *Speaking Personally* is almost the same, but the number of words differs significantly, due to the two temporal disruptions marked by a tilde (personal remarks and opinions by the narrator on general topics). *Speaking Personally* and *Word Game* stories are quite close but the former positions the last situations (parallel situations) in front of a building (the railway station) whereas the latter positions them somewhere on an urban square (*cour de Rome*).

6 Implementation and Evaluation

The approach has been tested on different types of text: a play (Sophocles), crime stories (A. Christie, B. Akunin), science fiction and fantasy (S. Lem, T. Pratchett, D. Adams), French literature (H. Balzac, R. Queneau), reports of interviews (*e.g.* ethnological research) or historical texts (*e.g.* 16th century textual building inventories).

The corpus includes textual content written in English, French and Polish. One of the reasons of this choice was to check that the approach is workable in different languages. Another reason was to test the impact of a given natural language - *i.e.* test if the segmentation of a given textual content, once translated into another language, remains fully consistent with the original.

As mentioned in Sect. 4 the annotation step results in alphanumeric codes associated to each situation. These codes, along with bibliographic data and other general information concerning the texts, are stored in an RDBMS structure. They are interpreted on the fly (Perl scripts) to produce SVG (Scalable Vector Graphic) interactive visualisations available inside standard web browsers.

6.1 Evaluation

An early "feasibility" evaluation was carried out with a group of non-experts (twelve students in mechanical engineering) in order to get a first feedback on the ***knowledge modelling bias*** (segmentation into sequences of situations). We asked testers to depict an everyday series of actions, such as their home to work routine, using the graphical codes. We then asked then to complement the description of each individual situation with one or several qualitative parameters of their choice. Some recurrent parameters emerged, such as sound, amount of light, mood, *etc*. What this evaluation procedure did usefully underline is that the logic behind the segmentation protocol is easily understood, and somewhat intuitive.

Yet there is a clear difference between asking testers to analyse one of their own everyday routine in terms of series of situations and having them uncover these situations from a textual content using predefined segmentation rules. In a second round we therefore implemented a more demanding evaluation setup, with this time eight testers from different countries (Marie Curie fellows focusing on reality-based 3D modelling – no native speakers of English) working on two extracts from novels written in English. It has to be said right away that the fact that testers were not native English speakers and the fact that the whole evaluation process (including the introduction to the approach, the reading of texts, the segmentation effort itself) lasted two hours and a half only undoubtedly minor the scope and significance of the evaluation and relativize the conclusions that can be drawn. Yet as will be discussed below the experience has impacted the way we can foresee future developments of the approach. It in particular helped us pinpoint a noticeable difference between perceptions of spatial disruptions (relatively consistent among readers) and temporal disruptions (very erratic). It also helped us understand that beyond what we considered at the beginning of this research as a key potential benefit of the approach (a systematic, discipline-independent segmentation and visualisation procedure) there is another, maybe even more promising benefit: helping different readers to *formalise* their own understanding of a narrative and thereby facilitating and structuring workgroup discussions.

The testers were first introduced to the approach, and shown the whole set of segmentation rules. Following, they were asked to work on a first text that they had to segment under supervision. This step was needed to make sure that the protocol was clear enough for them. These two phases lasted for an hour and a half. Testers were then left for one hour with a 1000 words text that they had to segment on their own, *i.e.*

on one hand they had to ***spot situation changes*** and on the other hand they had to ***qualify each situation*** with regards to space (what scale?) to time (any disruption?) actors (who is concerned?) and motion (do actors move?).

A central issue we wanted to raise was whether or not ***situations***, can easily and unambiguously be differentiated from one another. We analysed both the raw, quantitative results (number of situations spotted, types of scales identified, quantity of switches between static and dynamic situations, *etc.*) and the oral remarks made by the testers at the end of the evaluation.

Results show that generally speaking the concept of situation is quite easy to use – testers had no particular difficulty in spotting different situations and tagging them with values for the four parameters. But if the mechanism was found clear, we spotted a number of ambiguities deriving from two different issues: comprehension and individual interpretation of the segmentation rules on one hand, and inherent "fuzziness" of texts on the other hand. The open discussion that followed the gathering of the oral remarks was a particularly stimulating moment. Testers started comparing their own understandings of the text and explicit their interpretations –for example, how they understood spatial indications given in the text about a situation at a railway station and why they selected this or that scale from our theoretical model. In that example it turned out that testers used recollections, images of *one* specific railway station and based their choice on this specific example (Fig. 21).

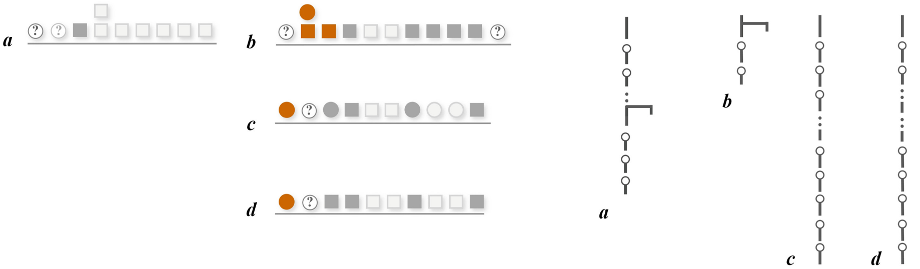


Fig. 21. A comparison between (a) results of the segmentation as we performed it and (b, c, d) results achieved by some testers, illustrated on the *spatial sequences* visualisation (left) and the *temporal continuity* visualisation (right). Differences in the *spatial sequences* visualisation are relatively limited in terms of number of situations spotted (9 *vs.* 10 *vs.* 11), but a little more significant in terms of scale identification: testers systematically tag one situation at least as corresponding to the urban scale, and several situations are tagged as "in a new space" (dark grey) when we considered space as unchanging in the text. Differences in the *temporal continuity* visualisation are by contrast quite sharp: testers understood, interpreted, temporal discontinuities in the flow of the narration in very different ways.

Comprehension Issues. The evaluation showed us that the testers had some difficulties with time discretisation and scale identification. Although testers globally understood the rules, they did not have enough time to get familiar with them before the test - they somehow discovered them as they progressed in the segmentation of the text.

We also noticed different individual interpretations of the segmentation rules: *e.g.* what does '*after an instant*' really mean? Testers disagreed on this very notion. What kind of space is '*a railway station'*? A building, a building and its surroundings, an inside, an outside? Here again each tester pictured what '*a railway station'* is his own way.

A certain number of segmentation rules as we had verbalised them turned out to be either too loosely defined, or too interpretative – typically the notion of ***parallel situations*** that encompasses someone spying on others from behind a window to a phone call connecting two people located in different parts of the world. One type of parallel situations appeared as particularly confusing, when several groups of people are in the same space but act independently of one another - in this case the rule itself needs rethinking.

Finally, some testers questioned the segmentation rules themselves when the rules, according to them, did not let them stick close enough to the text. In the text proposed *Kate* is driven to the airport in a taxi – but no mention is made of the taxi driver in the initial situation. Those testers considered that putative actors – here the taxi driver – should be mentioned, although according the segmentation rules they were given only actors mentioned in the text should be specified.

Briefly speaking, the evaluation showed that segmentation rules and definitions of scales, temporal disruptions, motion and actors need to be further clarified and illustrated by examples in order to pin down the concepts and reduce existing ambiguities. More generally the above comprehension issues clearly underline the fact that more time should be spent on explaining the segmentation rules prior to the evaluation itself. Moreover applying correctly the segmentation rules requires a thorough understanding of how space and time are discretised – which implies a steep learning curve.

At this stage the approach requires from readers and annotators a good understanding of the segmentation rules, but also keeping a certain "distance" with the text in order to avoid confusing what is really written, with what one may deduce, understand or imagine. What we asked the testers to do – segmenting of a text into an alphanumeric code using a set of segmentation rules and of discrete values – requires from annotators skills and capabilities. It definitely is a demanding task that limits the circle of people who can be expected to carry out the annotation step.

Inherent Fuzziness of Texts. There are a number of factors that impact the way space, time, actors and motion are verbalised by authors. Texts are written with a significant amount of *unsaid*, or *half-said* elements – voluntary omission of details, figures of speech, *etc.*

Consider this yet straightforward example: "*She set off in search of first a newspaper and then some coffee. She was then unable to find a working phone"*. Should the reader here consider time as continuous, or as interrupted for a short while, for a long while? The author does not say openly whether there is a time disruption or not. The same can happen when mentioning spaces, actors, or even motion. Texts are the way they are, and readers will anyway interpret and understand them differently, whatever semantic-based segmentation rules one may write – a feature of what A. Korzybski named *verbal levels* [25].

The evaluation showed the inherent fuzziness of texts can be seen as an obstacle, but also as a potential object of study, an opportunity for instance to use the segmentation rules in order to localise areas where readers interpret a text differently.

Interpreting the evaluation's results should however be done with caution. The segmentation's learning curve is definitely steep: further evaluation efforts are therefore clearly needed (for example finding a match between a text and a visualisation in a setup where several possibilities are shown). It also underlined unexpected potential benefits of the approach, in its current state of development:

- It helps comparing how different people understand and interpret the spatio-temporal content of a text.
- It enhances debate, and helps uncovering precisely (in the flow of the text) where alternative interpretations occur, and why.
- It facilitates the communication by one individual of his own understanding of a text by supporting (through visual means) his discourse on rhythms of a narrative in a context + focus manner;
- It could be used to weigh and compare the level of interpretation required from readers depending on the text or author.

7 Limitations and Perspectives

StorylineViz should be understood as a *proof-of-concept* study, aimed at developing a *generic* approach to narrative analysis, but with at start no other ambition than investigating to which extent coupling a specific segmentation bias (space/time/actors/motion) with visualisation solutions could renew the way we can understand, compare, debate textual content. Hence we do acknowledge this research's actual impact needs to be weighed with regards to a number of significant limitations, and notably the following:

- We consider that the corpus of texts used as test cases is representative in terms of variety, heterogeneity, but it definitely is a partial corpus.
- The evaluation phase should clearly be deepened – notably with regards to fine-tuned usage scenarios.
- The comprehensibility of the segmentation rules for a wider public should be better assessed, as well as the learning curve.
- The implementation is a robust one, but it certainly could be rethought or improved.

In addition to these general remarks, and anticipating potential perspectives of the research, two major challenges ahead need to be pinpointed: working on the end-user service (visualisations), working on the upstream process (segmentation and annotation of the texts).

7.1 Visual Reasoning: Still a Challenge

Our approach bases on the idea that the combination of a non-standard segmentation procedure with appropriate visualisations can offer users new opportunities to perform reasoning tasks, and uncover pieces of knowledge inside textual content. But such a statement can only be corroborated (or invalidated) if the experimental setup proposed to testers is fully satisfactory. The visualisations we ended on do show the idea is worth exploring, but the implementation is at this stage not fully satisfactory. For instance support for a visual cross-examination of texts "within the eyespan" [26] needs to be improved. Accordingly we consider that our study needs to be extended and deepened in order to state without doubt that the approach is indeed, generic, workable across various types of texts, and fruitful in terms of knowledge discovery.

7.2 The Impact of Manual Annotation

Even more significant in terms of limitation of the research's potential impact is the fact that the annotation process - *i.e.* the segmentation of texts – is to this day a manual process. This clearly undermines perspectives of application of such an approach on a large scale. But on the other hand it also opens a clear perspective (and challenge) for this research. The approach hits the limits of existing NLP based methods. Hence rounds of discussion we are at this stage having with VA and NLP partners to try and investigate how the approach could be developed on a large scale. Even if a fully automated annotation process would turn out to be out of reach, working on semi-automatic procedures in the context of the emergent crowdsourcing paradigm would clearly open tangible large-scale application perspectives. Furthermore, human annotation is by itself a meaningful activity, opening perspectives in terms of communication and comparative analysis of text interpretation. Both going towards more automation in the segmentation process, and sticking to a human process, can therefore be considered as lines of development of the approach.

8 Conclusion

In this contribution we present a research that aims at investigating how performing analyses of narratives (in a broad sense, encompassing texts that range from ethnological records to fictional stories) can be renewed by introducing an unusual segmentation bias and a series of ad hoc visualisations conveying insights on how the space, time, actors and motion components of the text interact in successive situations as the narrative unfolds. The main claim behind the approach is that extracting the spatio-temporal content of a narrative and visualising it as a series of situations can help spotting and exploring significant patterns, trends, exceptions across various types of texts.

In short, the main benefit expected from applying the approach can be summed up in two words: *knowledge discovery*. The corpus on which the approach has been tested remains partial, but the experimentation does show the approach is workable across various types of texts. Illustrated in this contribution on the specific case of R. Queneau's *Exercises in style*, it proved useful in uncovering unexpected and noteworthy patterns inside the 99 versions of the same series of events the book is made of.

The evaluation carried out paved the way towards usage scenarios that would focus more on assessing differences between reading and comprehension experiences than on the "automatization" of the segmentation process.

The visualisations produced until now show an interesting interpretative potential. They could be used for example to support teaching and learning activities, helping learners to quickly get a hold on patterns, trends, exceptions, and to carry out comparative analyses across texts (for instance using the approach in order to support pupils with learning disabilities such as *dyslexia*). We consider that, at this stage, the approach has proven workable, but will need further improvement loops (more case studies, more rounds of evaluation) before becoming fully operable.

References

1. ReNom: Navigating the works of Rabelais and Ronsard in search of people and places. http://renom.univ-tours.fr/en/project. Accessed 01 Mar 2017
2. Dudek, I., Blaise, J.Y.: StorylineViz: a [Space, Time, Actors, Motion]—segmentation method for visual text exploration. In: Fred, A., Aveiro, D., Dietz, J., Filipe, J., Bernardino, J., Liu, K. (eds.) Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, KDIR, vol. 1, pp. 21–32. SciTePress (2016). https://doi.org/10.5220/0006034600210032
3. Marshman, E., Van Bolderen, P.M.: Interlinguistic variation and lexical knowledge patterns. In: Madsen, B.N., Thomsen, H.E. (eds.) Managing Ontologies and Lexical Resources: Internationale Sprogstudier og Vidensteknologi, pp. 263–278. Litera, Copenhagen (2012)
4. Sabol, V.: Visual analysis of relatedness in dynamically changing repositories. In: MOVE-REALthematic School, Fréjus (2016)
5. Blaise, J.Y., Dudek, I.: Analyzing alternative scenarios of evolution in heritage architecture: modelling and visualization challenges. J. Multimed. Process. Technol. **3**(1), 29–48 (2012)
6. Blaise, J.Y., Dudek, I.: Spotting temporal co-occurrence patterns: the historySkyline visual metaphor. In: Proceedings of the 1st International Conference on Metrology for Archaeology, Benevento, pp. 378–383 (2015). ISBN: 978-88-940453-3-8
7. Matthew Paris Itinerary to the Holy Land (ca. 1250 A.D.). http://cartographic-images.net/Cartographic_Images/225.1_Palestine.html. Accessed 01 Mar 2017
8. Rosenberg, D., Grafton, A.: Cartographies of Time: A History of the Timeline. Architectural Press, Princeton (2012)
9. Yabuuchi, A.: Home to school diagram. In: Informational Diagram Collection, p. 215. Pie Books, Tokyo (2009)
10. Oelke, D.: Visual document analysis: towards a semantic analysis of large document collections. Ph.D. dissertation, University of Konstanz (2010)
11. Marazzato, R., Sparavigna, A.C.: Extracting Networks of Characters and Places from Written Works with CHAPLIN. CoRR - Computing Research Repository (2014). https://arxiv.org/abs/1402.4259
12. Bilenko, N.Y., Miyakawa, A.: Visualization of narrative structure analysis of sentiments and character interaction in fiction (2013). http://vis.berkeley.edu/courses/cs294-10-fa13/wiki/images/7/7b/AMNBpaper.pdf
13. Marshman, E., L'Homme, M.C., Surtees, V.: Verbal markers of cause-effect relations across corpora. In: Madsen, B.N., Thomsen, H.E. (eds.) Managing Ontologies and Lexical Resources: Internationale Sprogstudier og Vidensteknologi, pp. 159–174. Litera, Copenhagen (2008)
14. Spence, R.: Information Visualization. Pearson Addison-Wesley ACM Press, Harlow (2001)
15. Thomas, J.J., Cook, K.A.: Illuminating the path: the research and development agenda for visual analytics. IEEE Comput. Graph. Appl. **26**(1), 10–13 (2006)

16. Oelke, D., Spretke, D., Stoffel, A., Keim, D.: Visual readability analysis: how to make your writings easier to read. In: IEEE Symposium on Visual Analytics Science and Technology, pp. 123–130, VAST (2010)
17. Koch, S., John, M., Wörner, M., Müller, A., Ertl, T.: VarifocalReader—in-depth visual analysis of large text documents. IEEE Trans. Vis. Comput. Graph. **2**(12), 1723–1732 (2014)
18. Vuillemot, R., Clement, T., Plaisant, C., Kumar, A.: What's being said near 'Martha'? Exploring name entities in literary text collections. In: IEEE Symposium on Visual Analytics Science and Technology, pp. 107–114. VAST (2009)
19. Wanner, F., Fuchs, J., Oelke, D., Keim, D.A.: Are my children old enough to read these books? Age suitability analysis. Polibits: Res. J. Comput. Sci. Comput. Eng. Appl. **43**, 93–100 (2011)
20. Kergosien, E., Laval, B., Roche, M., Teisseire, M.: Are opinions expressed in land-use planning documents? Int. J. Geogr. Inf. Sci. **28**(4), 739–762 (2014)
21. Blaise, J.Y., Dudek, I.: Using abstraction levels in the visual exploitation of a knowledge acquisition process. In: Proceedings of I-Know 2005, Graz, pp. 543–552 (2005)
22. Blaise, J.Y., Dudek, I.: Profiling artefact changes: a methodological proposal for the classification and visualisation of architectural transformations. In: Digital Heritage, Proceedings of VSMM 2008—Virtual Systems and Multimedia, pp. 349–356. Archeolingua, Budapest (2008)
23. Aigner, W., Miksch, S., Schumann, H., Tominski, C.: Visualization of Time-Oriented Data. Human-Computer Interaction Series. Springer, London (2011). https://doi.org/10.1007/978-0-85729-079-3
24. Sabol, V.: Visualisation in the Web. http://kti.tugraz.at/sta-ff/vsabol/courses/mmis1/slides_vis.pdf
25. Korzybski, A.: The role of language in the perceptual processes. In: Blake, R., Ramsey, G. (eds.) Perception: An Approach To Personality, pp. 170–205. The Ronald Press Company, New York (1951)
26. Tufte, E.R.: Envisioning Information. Graphics Press, Cheshire (2001)

Multi-layer and Co-learning Systems for Semantic Textual Similarity, Semantic Relatedness and Recognizing Textual Entailment

Ngoc Phuoc An Vo(✉) and Octavian Popescu

IBM T. J. Watson Research Center, Yorktown Heights, USA
ngoc.phuoc.an.vo@ibm.com, o.popescu@us.ibm.com

Abstract. Similarity plays a central role in language understanding process. However, it is always difficult to precisely define on which type of data and what similarity metrics we can apply in order to assess the similarity of two texts. Previously, we proposed a four-layer system [69] that takes into account not only string and semantic word similarities, but also word alignment and sentence structure. Our system achieved new state of the art or competitive result to state of the art on different test corpora for the Semantic Textual Similarity (STS) task from 2012 to 2015. The multi-layer architecture helps to deal with heterogeneous corpora which may not have been generated by the same distribution nor same domain. In this extended paper, we looked into the correlation between the two semantic processing tasks Semantic Relatedness (a more broad task of STS) and Recognizing Textual Entailment (RTE) to construct a co-learning model where we integrated our multi-layer architecture and Corpus Patterns technique to ultimately improve the performances of both tasks.

Keywords: Machine learning · Natural Language Processing
Semantic Textual Similarity · Semantic Relatedness
Recognizing Textual Entailment · Corpus Patterns

1 Introduction

Exhaustive language models are difficult to build because overcoming the effect of data sparseness requires an unfeasible amount of training data. In the task of Semantic Text Similarity[1] (STS), the systems must quantifiably identify the degree of similarity between pairs of short pieces of text, like sentences. On the basis of relatively small training corpora, annotated with a semantic similarity score obtained by averaging the opinions of several annotators, an automatic system may learn to identify classes of sentences which could be treated in the same way, as their meaning is basically the same. It has been shown that good

[1] http://ixa2.si.ehu.es/stswiki/index.php/Main_Page.

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 54–77, 2019.
https://doi.org/10.1007/978-3-319-99701-8_3

results from STS systems may help to improve the accuracy on related tasks, such as Paraphrasing [20], Textual Entailment [11], Question Answering [65], etc.

However, building a system able to cope with various phenomena which fall under the umbrella of semantic similarity is far from trivial. Various types of knowledge must be considered when dealing with semantic similarity, and the methodology of linking together different pieces of information is a matter of research. It is almost always the case that the performances of a system do not vary consistently or predictably from corpora to corpora. The STS corpora used in STS competitions, and the task description papers [1–4] testify that there is no system that consistently scores the best across corpora, and big variation of system performance may occur.

The heterogeneity of sources considered for these corpora makes it difficult to maintain the hypothesis of the same probability distribution of terms for training and testing, therefore we have to adapt our system to handle this situation, which is better described as a mixture of more or less independent and unknown Gaussian.

This paper is extended from our previous paper [69]. From Chaps. 2 to 5, we reviewed the multi-layer system for Semantic Textual Similarity task (STS) which is modular having four principal layers: (i) string similarity, (ii) semantic word similarity, (iii) word alignment, and (iv) structural information. Our new contribution is from Chaps. 6 to 10 in which we analyzed the correlation between the two related semantic processing tasks: Semantic Relatedness and Recognizing Textual Entailment, and proposed a new co-learning model to improve the performance of these two tasks.

The paper continues as follows: in the next section we present the related works on semantic similarity which proved instrumental in the building of the actual system. In Sect. 3 we review the system based on four layers. Section 4 describe the experiment settings and Sect. 5 presents the evaluation results for the multi-layer system. In Sect. 6, we analyze the correlation between the Semantic Relatedness (SR) and Recognizing Textual Entailment (RTE) tasks, then Sects. 7 and 8 describe the corpus patterns approach for RTE. Sections 9 and 10 propose a co-learning system for SR and RTE, then the experiment to evaluated this system. The paper ends with conclusion section.

2 Related Work

The Semantic Text Similarity (STS) task has become one of the most popular research topics in NLP. Two main approaches have been widely used for tackling this task, namely Distributional Semantic Models (DSMs) and Knowledge-based similarity approaches.

Distributional Semantic Models (DSMs) is a family of approaches based on the distributional hypothesis [28], according to which the meaning of a word is determined by the set of textual contexts in which it appears. These models represent words as vectors that encode the patterns of co-occurrences of a word with other expressions extracted from a large corpus of language [56,66]. DSMs are very popular for tasks such as semantic similarity. The different meanings of a word are described in a space and words used in similar contexts are represented by vectors (near) in this space. On the basis of such methods, semantically similar words will appear in points near the (semantic) space. Textual contexts can be defined in different ways, thus giving rise to different semantic spaces.

Knowledge-based similarity approaches quantify the degree to which two words are semantically related using information drawn from semantic networks [14]. Most of the widely used measures (e.g. Leacock and Chodorow, Wu and Palmer, Lin, and Jiang and Conrath, among others) of this kind have been found to work well on the WordNet taxonomy. All these measures assume as input a pair of concepts, and return a value indicating their semantic similarity. Though these measures have been defined between concepts, they can be adapted into word-to-word similarity metrics by selecting for any given pair of words those two meanings that lead to the highest concept-to-concept similarity.

If we focus on sentence to sentence similarity, three prominent approaches are usually employed. The first approach uses the vector space model [40] in which each text is represented as a vector (bag-of-words). The similarity between two given texts is computed by different distance/angel measures, like cosine similarity, Euclidean, Jaccard, etc. The second approach assumes that if two sentences are semantically equivalent, we should be able to align their words or expressions. The alignment quality can serve as a similarity measure. This approach typically pairs words from two sentences by maximizing the summation of the word similarity of the resulting pairs [42]. The last approach employs different measures (like lexical, semantic and syntactic) from several resources as features to build machine learning models for training and testing [7,39,58,60,67].

As for the specific case of measuring semantic similarity between two given sentences, the Semantic Textual Similarity (STS) tasks[2], [3] have been officially organized and have received an increasing amount of attention [1–4].

The UKP (or also knows as DKPRO) [7] was the first-ranked system at STS 2012. This system used a log-linear regression model to combine multiple text similarity measures which range from simple measures (word n-grams or common subsequences) to complex ones (Explicit Semantic Analysis (ESA) vector comparisons [18], or word similarity using lexical-semantic resources). Beside this, it also used a lexical substitution system and statistical machine translation system to add additional lexemes for alleviating lexical gaps. The final models after the feature selection, consisted of 20 features, out of the possible 300+ features implemented.

By contrast, the best system at STS 2013, UMBC EBIQUITY-CORE [24], adopted and expanded the alignment approach into "align-and-penalize" by giving penalties to both the words that are poorly aligned and to the alignments causing semantic or syntactic contradictions. At the word level, it used a common

[2] http://www.cs.york.ac.uk/semeval-2012/task6/.

[3] http://ixa2.si.ehu.es/sts/.

Semantic Word Similarity model which is a combination of LSA word similarity and WordNet knowledge.

The DLS@CU [63] achieved best result at STS 2014. It used the word alignment approach described in the literature [62], which considered several semantic features, e.g. word similarity, contextual similarity, and alignment sequence. It [64] again achieved the best result as shown at STS 2015 using word alignment and similarities between compositional sentence vectors as its features. It adopted the 400-dimensional vectors developed in [8] using the word2vec toolkit [43] to extract these vectors from a large corpus (about 2.8 billion tokens). Word vectors between the two input sentences were not compared, but a vector representation of each input sentence was constructed using a simple vector composition scheme, then the cosine similarity between the two sentence vectors is computed as the second feature. The vector representing a sentence is the centroid (i.e., the componentwise average) of its content lemma vectors. Finally, these two features are combined using a ridge regression model implemented in scikit-learn [47]. Besides DLS@CU, it is very interesting that aligning words between sentences has been the most popular approach for other top participants ExBThemis [26], and Samsung [25].

Besides these approaches, a new semantic representation for lexical was proposed as semantic signature which is the multinomial distribution generated from the random walks over WordNet taxonomy where the set of seed nodes is the set of senses present in the item, [48]. This representation encompassed both when the item is itself a single sense and when the item is a sense-tagged sentence. This approach was evaluated on three different tasks Textual Similarity, Word Similarity and Sense Similarity; and it also achieved the state of the art on STS 2012 datasets.

Recently, more complex approaches have been proposed based on vector compositionality by means of operations, such as vector addition or tensor product among others. Earlier approaches implements the semantically compositional vector as the results of one of the operation from two independent vectors [15,44]. New efforts and approaches have been implemented to overcome this "independent" assumption of the compositional vector. The main innovations introduced by [21] and [9] is the introduction the co-occurrence vectors of observed composed constructions to train supervised composition models. [12] applies several compositional models, based on addition, multiplication, and recursive neural networks, on two tasks for text similarity: similarity judgments for short phrases [44] and on paraphrase detection based on Microsoft Research Paraphrase Corpus (MSRPC). Concerning the paraphrase detection task, [12] achieves good and promising results using less complex features than the best performing systems.

3 A Four-Layer System for STS

In this section we reviewed our four-layer system built from different linguistic features [69]. We constructed a pipeline system, in which each component produces different features independently and at the end, all features are consolidated by the machine learning tool WEKA, which learns a regression model for

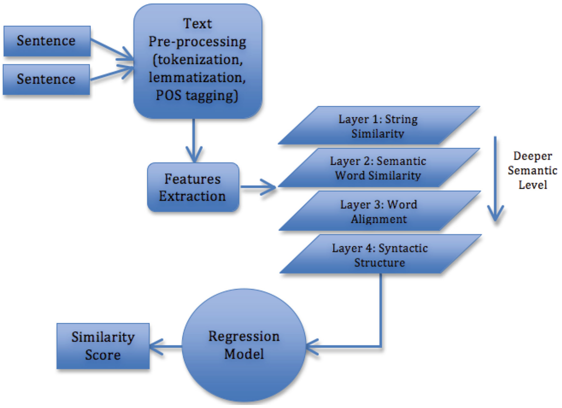

Sentence
Sentence
Text Pre-processing (tokenization, lemmatization, POS tagging)
Features Extraction
Layer 1: String Similarity
Layer 2: Semantic Word Similarity
Layer 3: Word Alignment
Layer 4: Syntactic Structure
Deeper Semantic Level
Regression Model
Similarity Score

Fig. 1. System overview [69].

predicting the similarity scores from given sentence-pairs. Beyond the typical features such as string similarity, character/word n-grams, and pairwise similarity, we add several distinguished features, such as syntactic structure information, word alignment and semantic word similarity. The System Overview in Fig. 1 [69] shows the logic and design processes in which different components connect and work together.

3.1 Data Pre-processing

The input data undergoes the data preprocessing in which we use Tree Tagger [59] to perform tokenization, lemmatization, and Part-of-Speech (POS) tagging. On the other hand, we use the Stanford Parser [33] to obtain the dependency parsing from given sentences.

3.2 Layer One: String Similarity

We use Longest Common Substring [23], Longest Common Subsequence [5] and Greedy String Tiling [73] measures.

Longest Common Substring is the longest string in common between two or more strings. Two given texts are considered similar if they are overlapping/covering each other (e.g. sentence 1 covers a part of sentence 2, or otherwise).

Longest Common Subsequence is the problem of finding the longest subsequence common to all sequences in a set of sequences (often just two sequences). It differs from problems of finding common substrings: unlike substrings, subsequences are not required to occupy consecutive positions within the original sequences.

Greedy String Tiling algorithm identifies the longest exact sequence of substrings from the text of the source document and returns the sequence as tiles (i.e., the sequence of substrings) from the source document and the suspicious document. This algorithm was implemented based on running Karp-Rabin matching [72].

3.3 Layer Two: Semantic Word Similarity

Semantic word similarity is the most basic semantic unit which is used for inferring the semantic textual similarity. There are several well-known approaches for computing the pairwise similarity, such as semantic measures using the semantic taxonomy WordNet [17] described by [29,31,35,36,55,74]; or other corpus-based approaches like Latent Semantic Analysis (LSA) [34], Explicit Semantic Analysis (ESA) [18], etc.

Among the approaches described above, we deploy three different approaches to compute the semantic word similarity: the pairwise similarity algorithm by Resnik [55] on WordNet [17], the vector space model Explicit Semantic Analysis (ESA) [18], and the Weighted Matrix Factorization (WMF) [22].

Resnik Algorithm returns a score denoting how similar two word senses are, based on the Information Content (IC) of the Least Common Subsumer (most specific ancestor node). As this similarity measure uses information content, the result is dependent on the corpus used to generate the information content and the specifics of how the information content was created.

Explicit Semantic Analysis (ESA) is a vectorial representation of text (individual words or entire documents) that uses a document corpus as a knowledge base. Specifically, in ESA, a word is represented as a column vector in the TF-IDF matrix [57] of the text corpus and a document (string of words) is represented as the centroid of the vectors representing its words. The ESA model is constructed by two lexical semantic resources Wikipedia and Wiktionary.[4,5]

Weighted Matrix Factorization (WMF) [22] is a dimension reduction model to extract nuanced and robust latent vectors for short texts/sentences. To overcome the sparsity problem in short texts/sentences (e.g. 10 words on average), the missing words, a feature that LSA/LDA typically overlooks, is explicitly modeled. We use the pipeline to compute the similarity score between texts.[6]

Besides these pairwise similarity methods, we also use the n-gram technique at character and word levels. We compare character n-grams [10] with the variance n = 2, 3, ..., 15. By contrast, we compare the word n-grams using the Jaccard coefficient done by Lyon [37] and containment measure [13] with the variance of n = 1, 2, 3, and 4.

3.4 Layer Three: Word Alignment

At the shallow level of comparing texts and computing their similarity score, we deploy two machine translation evaluation metrics: the METEOR [6] and

[4] http://en.wikipedia.org/wiki/Main_Page.

[5] http://en.wiktionary.org.

[6] http://www.cs.columbia.edu/~weiwei/code.html.

TERp [61]. However, our analysis shows that the TERp result does not really contribute to the overall performance, yet sometimes it may affect our system negatively. Hence, we remove this metric from the system.

Metric for Evaluation of Translation with Explicit ORdering (METEOR) [6] is an automatic metric for machine translation evaluation, which consists of two major components: a flexible monolingual word aligner and a scorer. For machine translation evaluation, hypothesis sentences are aligned to reference sentences. Alignments are then scored to produce sentence and corpus level scores. We use this word alignment feature to learn the similarity between words and phrases in two given texts in case of different orders.

3.5 Layer Four: Syntactic Structure

Intuitively, the syntactic structure plays an important role for the human being to understand the meaning of a given text. Thus, it also may help to identify the semantic equivalence between two given texts. We exploit the syntactic structure information by the mean of three different approaches: Syntactic Tree Kernel, Distributed Tree Kernel and Syntactic Generalization. We describe how each approach learns and extracts the syntactic structure information from texts to be used in our STS system.

Syntactic Tree Kernel. Given two trees T_1 and T_2, the functionality of tree kernels is to compare two tree structures by computing the number of common substructures between T1 and T2 without explicitly considering the whole fragment space. According to [45], there are three types of fragments described as the subtrees (STs), the subset trees (SSTs) and the partial trees (PTs). A subtree (ST) is a node and all its children, but terminals are not STs. A subset tree (SST) is a more general structure since its leaves need not be terminals. The SSTs satisfy the constraint that grammatical rules cannot be broken. When this constraint is relaxed, a more general form of substructures is obtained and defined as partial trees (PTs).

The Syntactic Tree Kernel is a tree kernels approach to learn the syntactic structure from syntactic parsing information, particularly, the Partial Tree (PT) kernel is proposed as a new convolution kernel to fully exploit dependency trees. The evaluation of the common PTs rooted in nodes n_1 and n_2 requires the selection of the shared child subsets of the two nodes, e.g. [S [DT JJ N]] and [S [DT N N]] have [S [N]] (2 times) and [S [DT N]] in common. We use the tool svm-light-tk 1.5 to learn the similarity of syntactic structure.[7]

Syntactic Generalization (SG). Given a pair of parse trees, the Syntactic Generalization (SG) [19] finds a set of maximal common subtrees. Though generalization operation is a formal operation on abstract trees, it yields semantics information from commonalities between sentences. Instead of only extracting common keywords from two sentences, the generalization operation produces a

[7] http://disi.unitn.it/moschitti/SIGIR-tutorial.htm.

syntactic expression. This expression maybe semantically interpreted as a common meaning held by both sentences. This syntactic parse tree generalization learns the semantic information differently from the kernel methods which compute a kernel function between data instances, whereas a kernel function is considered as a similarity measure.

SG uses least general generalization (also called anti-unification) [49] to anti-unify texts. Given two terms E_1 and E_2, it produces a more general one E that covers both rather than a more specific one as in unification. Term E is a generalization of E_1 and E_2 if there exist two substitutions σ_1 and σ_2 such that $\sigma_1(E) = E_1$ and $\sigma_2(E) = E_2$. The most specific generalization of E_1 and E_2 is called anti-unifier. Technically, two words of the same Part-of-Speech (POS) may have their generalization which is the same word with POS. If lemmas are different but POS is the same, POS stays in the result. If lemmas are the same but POS is different, lemma stays in the result. The software is available here.[8]

Distributed Tree Kernel (DTK). [75] This is a tree kernels method using a linear complexity algorithm to compute vectors for trees by embedding feature spaces of tree fragments in low-dimensional spaces. Then a recursive algorithm is proposed with linear complexity to compute reduced vectors for trees. The dot product among reduced vectors is used to approximate the original tree kernel when a vector composition function with specific ideal properties is used. The software is available here.[9]

4 Datasets and Experiment Settings

The STS datasets [1–4] are constructed from various sources associated with different domains, e.g. newswire headlines, paraphrases, video description, image captions, machine translation evaluation, Twitter news and messages, forum data, glosses combination of OntoNotes, FrameNet and WordNet, etc. Only in STS 2012, the train and test datasets are provided, since STS 2013 onward, no new training dataset is given, but only the new test dataset, whereas datasets in previous years can be used for training. Except the setup in STS 2012 where several of test sets have designated training data, the STS 2013, 2014 setups are similar to STS 2015 with no domain-dependent training data. This domain-independent character of STS data is a great challenge for any system to achieve consistent performance. The detail of datasets described in Table 1 [69].

5 Evaluations on STS

The results are obtained with Pearson correlation, which is the official measure used in both tasks.[10] The overall result is computed by the Weighted Mean of the Pearson correlations on individual datasets which is weighted according to the

[8] https://code.google.com/p/relevance-based-on-parse-trees.

[9] https://code.google.com/p/distributed-tree-kernels.

[10] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.

Table 1. Summary of STS datasets in years 2012–2015 [69].

Year	Dataset	#pairs	Source
2012	MSRpar	1500	newswire
2012	MSRvid	1500	video descriptions
2012	OnWN	750	OntoNotes, WordNet glosses
2012	SMTnews	750	Machine Translation evaluation
2012	SMTeuroparl	750	Machine Translation evaluation
2013	headlines	750	newswire headlines
2013	FNWN	189	FrameNet, WordNet glosses
2013	OnWN	561	OntoNotes, WordNet glosses
2013	SMT	750	Machine Translation evaluation
2014	headlines	750	newswire headlines
2014	OnWN	750	OntoNotes, WordNet glosses
2014	Deft-forum	450	forum posts
2014	Deft-news	300	news summary
2014	Images	750	image descriptions
2014	Tweet-news	750	tweet-news pairs
2015	Images	750	image descriptions
2015	headlines	750	newswire headlines
2015	answers-students	750	student answers
2015	answers-forum	375	forum answers
2015	belief	375	forum

Table 2. Evaluation results on STS 2012 datasets [69].

System	MSRpar	MSRvid	SMTeur	OnWN	SMTnews	Mean
Baseline	0.433	0.30	0.454	0.586	0.391	0.436
DKPro	0.62	0.808	0.376	0.657	0.462	0.584
UKP (1st)	0.683	0.874	0.528	0.664	0.494	0.677
Takelab (2nd)	0.734	0.880	0.477	0.680	0.399	0.675
SOFT-CARDINALITY (3rd)	0.641	0.856	0.515	0.711	0.483	0.671
ADW [48]	0.694	0.887	0.555	0.706	0.604	**0.711**
OurSystem (OS)	**0.748**	**0.894**	0.458	**0.755**	0.505	**0.711**

Table 3. Evaluation results on STS 2013 datasets [69].

System	FNWN	Headlines	OnWN	SMT	Mean
Baseline	0.215	0.540	0.283	0.286	0.364
DKPro	0.385	0.706	0.784	0.317	0.569
UMBC_EBIQUITY_PairingWords (1st)	0.582	0.764	0.753	0.380	**0.618**
UMBC_EBIQUITY_galactus (2nd)	0.743	0.705	0.544	0.371	0.593
deft-baseline (3rd)	0.653	0.843	0.508	0.327	0.580
OurSystem (OS)	0.450	0.732	**0.843**	0.356	**0.611**

Table 4. Evaluation results on STS 2014 datasets [69].

Systems	deft-forum	deft-news	Headlines	images	OnWN	tweet-news	Mean
Baseline	0.353	0.596	0.510	0.513	0.406	0.654	0.507
DKPro	0.452	0.713	0.697	0.777	0.819	0.722	0.714
DLS@CU (1st)	0.483	0.766	0.765	0.821	0.859	0.764	0.761
MeerkatMafia (2nd)	0.471	0.763	0.760	0.801	0.875	0.779	0.761
NTNU (3rd)	0.531	0.781	0.784	0.834	0.850	0.676	0.755
OurSystem (OS)	0.508	0.762	0.765	0.818	**0.896**	0.749	**0.768**

Table 5. Evaluation results on STS 2015 datasets [69].

System	ans-forums	ans-students	belief	headlines	images	Mean
Baseline	0.445	0.665	0.652	0.531	0.604	0.587
DKPro	0.696	0.712	0.699	0.766	0.808	0.746
DLS@CU-S1 (1st)	0.739	0.773	0.749	0.825	0.864	**0.802**
ExBThemis-themisexp (2nd)	0.695	0.778	0.748	0.825	0.853	0.794
DLS@CU-S2 (3rd)	0.724	0.757	0.722	0.825	0.863	0.792
OurSystem (OS)	0.713	0.744	0.733	0.808	0.858	**0.783**

Table 6. Comparison on all STS datasets [69].

Settings	2012	2013	2014	2015
Gain/Baseline	0.275	0.247	0.261	0.196
Gain/DKPro	0.127	0.042	0.054	0.037
Dist2SOTA	0.034	−0.007	0.007	−0.019

number of sentence pairs in that dataset. We compare our system's performance with the baseline and the top three systems in each STS competition in years 2012, 2013, 2014 and 2015.

Performance Comparison on all STS Datasets. Tables 2, 3, 4, and 5 [69] show our system performance in each year. In overall, Table 6 shows the side-by-side comparison between our system and the baseline, the DKPro and the state-of-the-art (SOTA) systems on all STS datasets. This confirms our stable and consistent performance which always overcomes the baseline (large margin 20–27%) and DKPro (4–13%), and achieves better or competitive results to SOTA systems.

Comparison to DKPro. Table 6 [69] shows that though we adopt some string and word similarity features from DKPro, our system always outperforms DKPro. The main difference between our system and DKPro is that by adding two important modules of processing word alignment and syntactic structure, we consider more linguistic aspects in semantic inference leading to more robust and comprehensive capability to compute the semantic similarity. This proves that

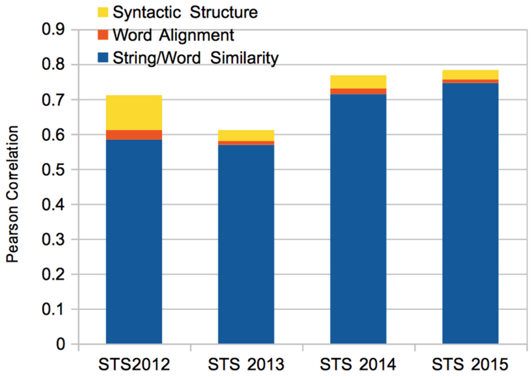


Fig. 2. Component analysis [69].

this approach of multi-layer infrastructure optimizes the system performance by delegating and capturing various linguistic phenomena by proper semantic layers, leading to higher precision and correlation.

Component Analysis. Figure 2 [69] presents the analysis for each individual component in our STS system. It shows the significance of each layer into the overall performance on STS 2012, 2013, 2014 and 2015 datasets. Despite the fact that string and word similarity layer occupies a larger portion in the overall performance, the significance of other semantic layers is undenied. The design of multi-layer system improves the overall performance from 3.7–12.7% more by better robustness and comprehension to handle more complicated semantic information via deeper semantic layers.

Accordingly, we can claim that our system consistently and stably performs at the state of the art or top-tier level on all STS datasets from 2012 to 2015. The framework of four different semantic layers helps our system handle heterogeneous data from STS successfully. By delegating and assigning different semantic layers which deal with different types of information, the system can cope with and adapt to any unknown domain data. This hypothesis is proven by the constant performance on various datasets derived from different domains in STS.

6 Correlation Between the Semantic Relatedness and Textual Entailment

As understanding sentence's meaning is a crucial challenge for any computational semantic system, there are several attempts on creating corpora for evaluating this task, including Recognizing Textual Entailment (RTE) and Semantic Textual Similarity (STS) or Relatedness (SR).

The RTE task requires the identification of a directional relation between a pair of text fragments, namely a text (T) and a hypothesis (H). The relation (T → H) holds if, typically, a human reading T would infer that H is most likely true.

In contrast, the SR task which is very much similar to the STS task requires to identify the degree of relatedness that exists between two text fragments (phrases, sentences, paragraphs, etc.), where similarity is a broad concept and its value is normally obtained by averaging the opinion of several annotators. While the concept of semantic similarity is more specific and it only includes the "is a" relations between two texts, the semantic relatedness may be broader and may include any relation between two terms such as antonymy and meronymy.

The Sentences Involving Compositional Knowledge (SICK) corpus the first dataset which contains the manual annotation for both tasks Semantic Relatedness (SR) and Textual Entailment (TE) [38]. It was created to evaluate the Compositional Distributional Semantic Models (CDSMs) handling the challenging phenomena, such as contextual synonymy and other lexical variation phenomena, active/passive and other syntactic alternations, impact of negation, quantifiers and other grammatical elements, etc. The SICK includes a large number of sentence pairs (around 10,000 English sentence pairs) that are rich in the lexical, syntactic and semantic phenomena that CDSMs are expected to account for, but it is not required to deal with other aspects of existing datasets (multiword expressions, named entities, numbers,) that are not within the domain of compositional distributional semantics. Each sentence pair in SICK is annotated for semantic relatedness score in the semantic scale [1–5] and textual entailment relation in 3-way: ENTAILMENT, NEUTRAL, and CONTRADICTION.

The literature [68] shows that there is a correlation between the SR scores and TE relations in which a certain TE value implies a SR score within a given interval. We adopt this idea to develop a system for enhancing the accuracy for both tasks.

7 Corpus Patterns for Semantic Processing

In this section we present a methodology for acquiring corpus patterns (CP) for target verbs. The patterns encode the lexical, syntactic and semantic information needed for meaning processing. Recognizing such patterns in text possibly draws sound inferences regarding the similarity or entailment of a pair of verbal phrases. Firstly, we analyze a property of natural languages, namely nonambiguity phrase, which leads to the definition of CP, then we review the CPs acquisition methodology.

7.1 Ambiguous and Non-ambiguous Phrases

Some phrases, just like words, are ambiguous and the senses of their words change according to context. For example, the phrase *I see* is ambiguous, as at least the sense of the verb *see* is not clear; it could mean *understand* or *perceive things*. However, by considering a larger context, the phrase may become unambiguous, in which, the senses of the words in the phrases do not change anymore whatever new context is added to the left or right. We call such phrases as sense stable phrases (SSPs). For example, by considering a larger context for *I see*, as in *I see your dog* or as in *I see your reason*, we obtain a different SSPs.

We are interested especially in minimal SSP, which is the minimally necessary context around the verb that creates a SSP. We show that words inside a SSPs are characterized by specific relationships, which combine lexical and syntactic information. On the basis of these relationships, we can derive an automatic methodology for acquisition and recognition SSPs. Minimal SSPs correspond to a pattern that combines syntactic roles and ontological traits, called lexical or semantic types [27,53] of the words. For example:

– sbj=[Human] see#1 obj=[PhysicalEntity]
– sbj=[Human] see#3 obj=[Stating]

– sbj=[Human] drive#1 prepTO=[Location]
– sbj=[Human] drive#5 prepTO=[PsychoState]

where sbj, obj, prepTO mark the syntactic function, verbs at #1, #3 and #5 mark the verb sense according to WordNet, and between the square brackets we mark the semantic types as specified in an ontology. We call the above patterns as Corpus Patterns (CP) because they are derived entirely on the basis of the verb behavior in corpus. Each context matched by a CP is a SSP, and it means that the necessary meaning information is encoded in the CP to remain the sense of the words in any other contexts.

Due to the above relationships between the senses of the words in a minimal SSP, in order to process the meaning of the verbal phrase, we need to find the semantic type of just one of the elements of the corresponding CP. In next section, we use this property to resolve semantic tasks.

7.2 Acquisition of Corpus Patterns

The supervised techniques to cluster corpus examples to obtain candidates of CPs is described in [32,53]. The basic idea is to cluster corpus examples of verb phrases according to their syntactic similarity. Then, we cluster the words on the same syntactic position according to their lexical property and the verb sense. The same class of words is represented by a semantic type that is taken from an ontology. We use the SUMO ontology [46], which is an ontology aligned with WordNet, on examples taken from SemCor [41].

Another similar approach is to use a statistical approach to extract the CPs from other resources, such as Pdev [30] or OntoNotes [71], via stochastic finite automata or using Naive Bayesian formula [50,54]. Basically, for each class of verbs defined in Pdev or VerbNet, we construct a confusion matrix for each syntactic slot and SUMO attribute. The posterior probability of a CP is computed from the priors in the confusion matrix. The process is sound, because the CPs have a regular language form, thus there is a finite number of differences between the CPs that are associated with a certain verb. The output is a list of corpus patterns which use WordNet sense and SUMO semantic types. An example of the Corpus Pattern acquisition is described in Fig. 3 [68].

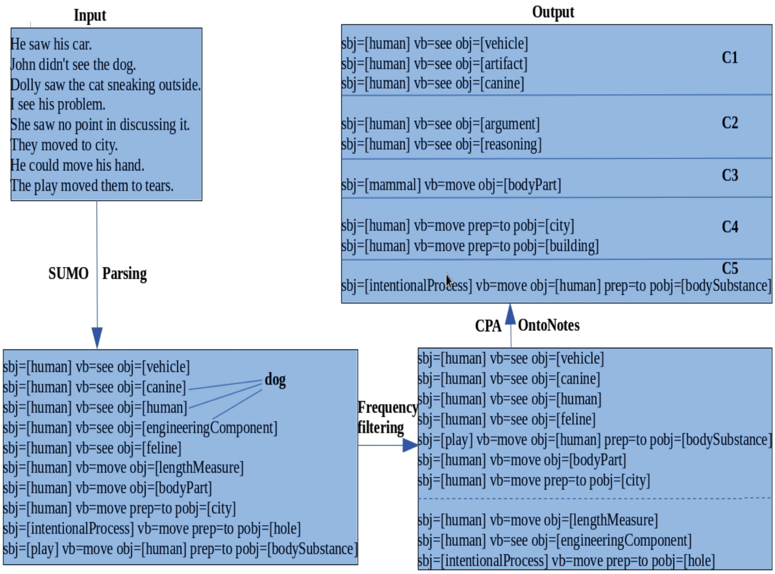

Input
He saw his car.
John didn't see the dog.
Dolly saw the cat sneaking outside.
I see his problem.
She saw no point in discussing it.
They moved to city.
He could move his hand.
The play moved them to tears.
SUMO
Parsing
sbj=[human] vb=see obj=[vehicle]
sbj=[human] vb=see obj=[canine]
sbj=[human] vb=see obj=[human]
dog
sbj=[human] vb=see obj=[engineeringComponent]
sbj=[human] vb=see obj=[feline]
sbj=[human] vb=move obj=[lengthMeasure]
sbj=[human] vb=move obj=[bodyPart]
sbj=[human] vb=move prep=to pobj=[city]
sbj=[intentionalProcess] vb=move prep=to pobj=[hole]
sbj=[play] vb=move obj=[human] prep=to pobj=[bodySubstance]
Frequency
filtering
sbj=[human] vb=see obj=[vehicle]
sbj=[human] vb=see obj=[canine]
sbj=[human] vb=see obj=[human]
sbj=[human] vb=see obj=[feline]
sbj=[play] vb=move obj=[human] prep=to pobj=[bodySubstance]
sbj=[human] vb=move obj=[bodyPart]
sbj=[human] vb=move prep=to pobj=[city]
sbj=[human] vb=move obj=[lengthMeasure]
sbj=[human] vb=see obj=[engineeringComponent]
sbj=[intentionalProcess] vb=move prep=to pobj=[hole]
CPA
OntoNotes
Output
sbj=[human] vb=see obj=[vehicle]
sbj=[human] vb=see obj=[artifact]
sbj=[human] vb=see obj=[canine]
C1
sbj=[human] vb=see obj=[argument]
sbj=[human] vb=see obj=[reasoning]
C2
sbj=[mammal] vb=move obj=[bodyPart]
C3
sbj=[human] vb=move prep=to pobj=[city]
sbj=[human] vb=move prep=to pobj=[building]
C4
C5
sbj=[intentionalProcess] vb=move obj=[human] prep=to pobj=[bodySubstance]

Fig. 3. Acquisition of Corpus Patterns.

8 Recognizing Textual Entailment via Corpus Patterns

The literature [52] shows a direct relation between corpus patterns and textual entailment. CPs represent an intermediate level of information representation between text and logical formula, and they encode the necessary contextual information for deciding on the phrase meaning. Thus, they can be used as input for a logical decision process. By matching CPs against a pair of sentences, we can decide the correct entailment relationship between them. We can manage the relationship between a pair of sentences with different polarity, because CPs can handle the differences between positive vs. negative sentences entailment. It is essentially the same set of conditions, except for the fact that in a negative sentence, the scope of negation is firstly determined and then the entailment decision is inverted according to the negation logic.

8.1 Positive Sentences Entailment via Corpus Patterns

Considering the following examples in the corpus:

- (s1) *A lemur is biting a person's finger.*
- (s2) *An animal is biting a person's finger.*

This pair of sentences has the same corpus pattern matched, so we can decide that they have Entailment relation as same as the manual annotation. This means that the same sense of the verb is used in both sentences. It is sufficient to observe individual relations between the slots of the CPs, because CPs ensure strict entailment conditions. Let's say CP_1 matching the s1 and CP_2 matching the s2, so the sentence s1 entails the sentence s2 if the semantic type (ST) present in a slot of CP_1 is a hyponym of the semantic type of CP_2. In general, we have the following schema:

CP_1: [sbj=ST_1 $\mathrm{v}_1\#n_1$ obj=ST_2 pp=ST_3] → CP_2: [sbj=ST_4 $\mathrm{v}_2\#n_2$ obj=ST_5 pp=ST_6], if:

- $\mathrm{v}_1\#n_1 \rightarrow \mathrm{v}_2\#n_2$, and
- $\mathrm{ST}_1 \rightarrow \mathrm{ST}_4$ (ST_1 is a hyponym of ST_4), and
- $\mathrm{ST}_2 \rightarrow \mathrm{ST}_5$ (ST_2 is a hyponym of ST_5), and
- $\mathrm{ST}_3 \rightarrow \mathrm{ST}_6$ (ST_3 is a hyponym of ST_6).

In the example above, *lemur* is a hyponym of *animal*, thus we have entailment. The condition $\mathrm{v}_1\#\mathrm{n}_1 \rightarrow \mathrm{v}_2\#\mathrm{n}_2$ requires that the two main verbs must belong to the same WorNet synset, or VerbNet group, or they are synonyms (see Sect. 8.5). The hyponym condition is rather strict, as there is no entailment if the words are characterized by the same semantic type, or in the reverse order.

8.2 Negative Sentences Entailment via CP

If a sentence without containing negation elements is matched by a CP, we can infer that there is at least an entity of each semantic type present in the CP. The sentence (s3): "*Two children are hanging on a large branch*" is matched by the CP *[Human] hang_on#1 [PhysicalObject]*, which implies that there are two entities of type *[Human]* for which a fact is asserted. Thus, a sentence that negates the existence of the two entities, or negates the fact asserted about the two entities, is in a contradiction relationship with the first sentence. Negation is realized by negative markers, such as "*there is no*" for existential negation, and "*not*" for factual negation. For example, the corresponding sentence of (s3) is (s4): "*There are no children hanging on a large branch*".

Both (s3) and (s4) are matched by the same pattern, except that sentence 4 contains a negative marker "*there are no*". Thus, the entailment decision should be reverse, which is Contradiction. In general, the corpus pattern conditions to be checked in case of negative polarity are:

1. check if there is existential or factual negation marker
2. remove the negation marker
3. check the CP conditions for positive sentences
4. if positive, then the first sentence contradicts the second.

8.3 Linguistic Pre-processing

The corpus pattern methodology described in the previous two sections considers the head of the nominal phrases of sentences in - form. However, there are different linguistic forms which require processing for accurate corpus pattern matching. We have considered in our system two main linguistic levels for pre-processing: syntactic and lexical. At the syntactic level, a sentence may exhibit different forms such as passive, coordination, negation, apposition, relative clauses, etc. At the lexical level, the constituents of a sentence may display synonymic and antonymic phrases, adverbial and adjectival modifiers, metonymy, synecdoche, semantic void heads, etc. We considered that the scope of negation and of coordination is the whole phrase, this hypothesis is true in more than 99% of the time in the SICK corpus.

8.4 Syntactic Pre-processing

Passive. The Stanford parser indicates explicitly the passive form, so we simply map the agent and passive subject to the corresponding subject and direct object of the active form.

Coordination. From the output of the Stanford parser, we can identify coordination and consider the cases when the subject, verb or object is multiple. We create a set of corpus patterns for each element of the coordination. For example, the sentence "*A man, a woman and two girls are walking on the beach*" is matched by the pattern *sbj=[Human] walk prepON=[Location]* and we keep a list of three different instances of this pattern with *sbj=[Human]* instantiated by *man, woman and girl* respectively.

Each of the pattern instances are checked for entailment against the pattern of the second sentence in the pair, according to the procedure described in Subsect. 8.1.

Negation. The two types of negation, existential and factual, are resolved according to the rules in Sect. 8.2. The scope of negation is considered to be the whole constituent, the subject for the negative existential marker, "*there is no*", and the verbal phrase for the negative factual marker, "*not*".

Apposition or Relative Clause. These syntactic constructions bring extra information to the head of the main constituent. This information is recorded for the head and used to decide whether the corresponding slots in the CPs observe the entailment conditions as described in Sect. 8.1.

8.5 Lexical Pre-processing

As the lexical properties of words are important and need to be learned, we used the SICK training data to learn the lexical features involved in entailment decision. The learning procedure works as follows:

1. Consider all the pairs of sentences which have a relatedness score above or equal 4 (highly similar) and annotated with Entailment or Contradiction relation.
2. Identify the CPs for both sentences.
3. if the verb is different or if the elements of the pattern are different, then learn that the respective phrases are synonymic or antonymic, according to the entailment value, entail or contradiction, respectively.

Supervised Lexical Learning. Using the technique above, we extracted three types of phrases: synonymic, antonymic and semantically void phrases. The output was:

- **942 pairs of synonymic phrases.** The list contains also metonymic or synecdochtic pairs, like "*field*" vs. "*grass*", but we do not know how many of them there are.
- **47 antonymic phrases**, like "*empty*" vs. "*fill*", or "*climb*" vs. "*get down*", which trigger contradiction between the sentences with no negative markers.
- **6 semantically void phrases**, such as "*a group of people*". These types of expressions play an important role in the entailment decision, because instead of the head of such expressions, like "*group*", we need to consider the prepositional modifier, "*people*" in order to have a correct CP match.

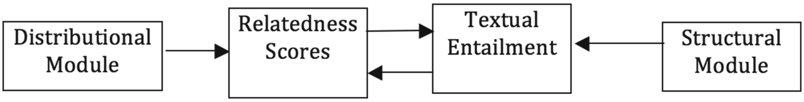


Fig. 4. SR revision scores architecture if the entailment test is positive/neutral then a quantity is added/subtracted to the SR score such that the correlation between TE value and SR score is maximized. The above parameters are set at training.

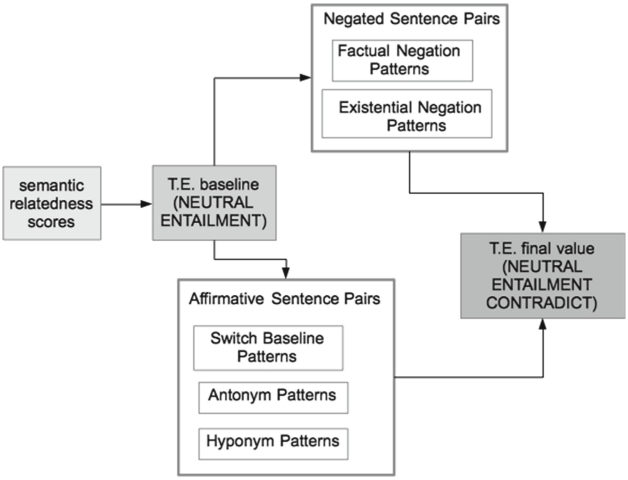

Negated Sentence Pairs
Factual Negation Patterns
Existential Negation Patterns
semantic relatedness scores
T.E. baseline (NEUTRAL ENTAILMENT)
T.E. final value (NEUTRAL ENTAILMENT CONTRADICT)
Affirmative Sentence Pairs
Switch Baseline Patterns
Antonym Patterns
Hyponym Patterns

Fig. 5. Structural module.

9 The Co-learning System for SR and TE

The mutual relationship suggests that a dual leverage architecture may be beneficial in addressing the ST and TE tasks. The main idea is to be able to adjust the TE or SR final decision taking into consideration an initial estimate. In general, good prediction can be obtained for SR by using a distributional strategy. Words are aligned between the two sentences by their similarity and the pair SR score is computed on the basis of individual similarities penalized accordingly to the alignment schema. To decide on the TE values, usually a more local and structural analysis focusing on various linguistics aspects - coordination, negation, semantic roles, etc. must be considered as well. However adding this complexity comes with a price on accuracy. Therefore is better to have a more robust way to ensure that the structural analysis does not deviate due to errors.

We propose an architecture in which the first SR score by the system described in Sect. 3 is used to estimate the entailment. A different module carries out a deep analysis on the sentence structure based on corpus patterns in order to decide the entailment especially for those border line cases signaled by the SR score, that is scores that are not very high but not very low either The entailment judgment is used to recompute the SR score, according to the following rule described in Fig. 4 [68].

However, the Structural Module in Fig. 5 [68] should employ a different class of techniques for determining the entailment. As the name of this module shows, a structural analysis of the sentence has to be performed. A technique that uses structural information to infer the entailment relationship was presented in [51,70]. The structural module should involve more detailed analyses of the semantics of the sentence. Then the initial SR scores are recomputed after the entailment decision is made.

Table 7. Performance analysis.

sys	*passive*		*negation*		*coordination*		*relative*		*synonymy*		*antonymy*		*void*	
	SR	TE	SR	TE	SR	TE	SR	TE	SR	TE	SR	TE	SR	TE
DKPRO	81	67	81	79	**79**	61	80	59	65	41	43	37	78	69
DKPRO_TTK	66	47	71	58	60	47	67	44	41	45	40	33	72	65
ML	86	69	**87**	81	78	60	80	62	65	44	41	38	82	**73**
ML+CP	**87**	**92**	81	**90**	67	**62**	**82**	**75**	71	**79**	**69**	**73**	**81**	71
DTK	62	47	74	68	57	45	58	46	63	43	42	33	70	65
SG	67	53	71	73	60	56	66	53	71	52	44	33	72	65
STK	70	58	73	70	68	61	71	61	**72**	67	50	38	71	**73**

10 Experiments

In this section, we analyze the performance of the above systems on the SICK test corpus. We present the contributions of each of the procedures presented in the previous sections and we compare them against the official baseline, against the distributional module, and against three tree kernel approaches, in order to calculate precisely the improvement brought by considering structural information via corpus patterns. The tree kernel approaches are reported to perform efficiently and effectively on processing syntactic trees using three proposed approaches Syntactic Tree Kernel (STK) [45], Syntactic Generalization (SG) [19] and Distributed Tree Kernel (DTK) [75]. All these systems have been trained on the SICK training corpora. For the STK system we also use the MSR paraphrasing corpus [16] in order to enhance the system capability to deal with synonymic phrases. In order to understand the behavior of ML which is our multi-layer system proposed in Sect. 3 and ML+CP which is the multi-layer system plus the corpus patterns described in Sect. 9, we also present the results using DKPRO system, and DKPRO combining with three kernel approaches DKPRO_TTK.

We start by analyzing the performances of the above systems for each type of sentences presented in Sect. 8.3, see Table 7. For SR score the Pearson correlation is reported, and for TE relation the accuracy is reported. The exact real number of pairs in each subset is not known, these sets are compiled by our recognizing pre-processing procedures based on the Stanford parser output. However, a manual sample check suggests that these estimates are pretty good, less than 4% errors in recognizing these cases.

From Table 7 we learn that structural information really helps in improving the accuracy, but it needs to take into account also the semantic types. The tree kernel approaches performed poorly by themselves. However, the pairs involving synonymic and antonymic knowledge are difficult for all systems. For semantic void all systems performed notably better. Apparently simply ignoring this constructions improves the results as the distributional models outperformed the structural ones. However, we think this is due to the fact that the corpus is biased, in the sense that the sentences with semantic void elements have many other words in common and are usually in entailment relation, thus a system may look, incorrectly, only to the distributional clues. Another salient thing is the fact that combining DKPRO and three kernel approaches into a single classifier does not look like a good idea, as this system scores even with 8% lower than individual three kernel methods.

In Table 8, BST stands for the best system on each task in Semeval 2014. The ML+CP improves the performance of ML, and outperforms the BST on both tasks.

Table 8. SICK corpus global results.

System	SR	TE	*System*	SR	TE
BST	82.8	84.6	DKPRO+CP	76.3	78.1
ML	83.1	81.1	DTK	69.5	48.9
ML+CP	**86.6**	**87.3**	STK	72.3	56.6
DKPRO	69.3	48.9	SG	69.4	48.5

The results of these experiments show that the co-learning system based on the mutual relation between SR and TE is beneficial for all types of systems, for example the improvement obtained by DKPRO+CP vs. DKPRO.

11 Conclusion

In this paper, firstly, we reviewed our multi-layer architecture towards building a model to solve the greatest challenge of domain-independent data for Semantic Textual Similarity task. In this architecture, we unified the task into four main layers of processing to exploit the semantic similarity information from different presentation levels to overcome the variance of system's performance on data derived from various sources. As a result, we built a system that achieved state of the art or competitive result to state of the art on different corpora built from different domains for Semantic Textual Similarity task from 2012 to 2015. Secondly, we extended our work by analyzing the correlation between the two related semantic processing tasks Semantic Relatedness and Recognizing Textual Entailment, then we deployed and integrated our multi-layer architecture with Corpus Patterns techniques to build a co-learning model in which we successfully improved the performance of these both tasks for the SICK dataset.

References

1. Agirre, E., et al.: SemEval-2015 task 2: semantic textual similarity, English, Spanish and pilot on interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics, Denver (2015)
2. Agirre, E., et al.: Semeval-2014 task 10: multilingual semantic textual similarity. In: SemEval 2014, p. 81 (2014)
3. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W.: SEM 2013 shared task: semantic textual similarity, including a pilot on typed-similarity. In: In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics. Citeseer (2013)
4. Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A.: Semeval-2012 task 6: a pilot on semantic textual similarity. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 385–393. Association for Computational Linguistics (2012)
5. Allison, L., Dix, T.I.: A bit-string longest-common-subsequence algorithm. Inf. Process. Lett. **23**(5), 305–310 (1986)
6. Banerjee, S., Lavie, A.: Meteor: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)

7. Bär, D., Biemann, C., Gurevych, I., Zesch, T.: UKP: Computing semantic textual similarity by combining multiple content similarity measures. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 435–440. Association for Computational Linguistics (2012)
8. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: ACL, vol. 1, pp. 238–247 (2014)
9. Baroni, M., Zamparelli, R.: Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, pp. 1183–1193. Association for Computational Linguistics, Stroudsburg (2010). http://dl.acm.org/citation.cfm?id=1870658.1870773
10. Barrón-Cedeno, A., Rosso, P., Agirre, E., Labaka, G.: Plagiarism detection across distant language pairs. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 37–45. Association for Computational Linguistics (2010)
11. Berant, J., Dagan, I., Goldberger, J.: Learning entailment relations by global graph structure optimization. Comput. Linguist. **38**(1), 73–111 (2012)
12. Blacoe, W., Lapata, M.: A comparison of vector-based representations for semantic composition. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 546–556. Association for Computational Linguistics, Jeju Island (2012). http://www.aclweb.org/anthology/D12-1050
13. Broder, A.Z.: On the resemblance and containment of documents. In: Proceedings of Compression and Complexity of Sequences 1997, pp. 21–29. IEEE (1997)
14. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–47 (2006). https://doi.org/10.1162/coli.2006.32.1.13
15. Clark, S., Pulman, S.: Combining symbolic and distributional models of meaning. In: AAAI Spring Symposium: Quantum Interaction, pp. 52–55 (2007)
16. Dolan, B., Brockett, C., Quirk, C.: Microsoft research paraphrase corpus (2005). Accessed 29 Mar 2008
17. Fellbaum, C.: WordNet. Wiley Online Library, New York (1998)
18. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI, vol. 7, pp. 1606–1611 (2007)
19. Galitsky, B.: Machine learning of syntactic parse trees for search and classification of text. Eng. Appl. Artif. Intell. **26**(3), 1072–1091 (2013)
20. Glickman, O., Dagan, I.: Acquiring lexical paraphrases from a single corpus. In: Recent Advances in Natural Language Processing III, pp. 81–90. John Benjamins Publishing, Amsterdam (2004)
21. Guevara, E.: A regression model of adjective-noun compositionality in distributional semantics. In: Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics, GEMS 2010, pp. 33–37. Association for Computational Linguistics, Stroudsburg (2010). http://dl.acm.org/citation.cfm?id=1870516.1870521
22. Guo, W., Diab, M.: Modeling sentences in the latent space. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp. 864–872. Association for Computational Linguistics (2012)

23. Gusfield, D.: Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge University Press, Cambridge (1997)
24. Han, L., Kashyap, A., Finin, T., Mayfield, J., Weese, J.: UMBC ebiquity-core: semantic textual similarity systems. In: In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics (2013)
25. Han, L., Martineau, J., Cheng, D., Thomas, C.: Samsung: align-and-differentiate approach to semantic textual similarity. In: SemEval-2015, p. 172 (2015)
26. Hänig, C., Remus, R., De La Puente, X.: ExB themis: extensive feature extraction from word alignments for semantic textual similarity. In: SemEval-2015, p. 264 (2015)
27. Hanks, P., Pustejovsky, J.: A pattern dictionary for natural language processing. Revue française de linguistique appliquée **10**(2), 63–82 (2005)
28. Harris, Z.S.: Mathematical Structures of Language. Interscience Publishers, Geneva (1968)
29. Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. In: WordNet: An Electronic Lexical Database, vol. 305, pp. 305–332 (1998)
30. Jezek, E., Hanks, P.: What lexical sets tell us about conceptual categories. Lexis **4**(7), 22 (2010)
31. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008 (1997)
32. Kawahara, D., Peterson, D.W., Popescu, O., Palmer, M.: Inducing example-based semantic frames from a massive amount of verb uses. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (2014)
33. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, pp. 423–430. Association for Computational Linguistics (2003)
34. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. **25**(2–3), 259–284 (1998)
35. Leacock, C., Miller, G.A., Chodorow, M.: Using corpus statistics and wordnet relations for sense identification. Comput. Linguist. **24**(1), 147–165 (1998)
36. Lin, D.: An information-theoretic definition of similarity. In: ICML, vol. 98, pp. 296–304 (1998)
37. Lyon, C., Malcolm, J., Dickerson, B.: Detecting short passages of similar text in large document collections. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, pp. 118–125 (2001)
38. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A SICK cure for the evaluation of compositional distributional semantic models. In: Proceedings of LREC 2014, Reykjavik (Iceland): ELRA (2014)
39. Marsi, E., Moen, H., Bungum, L., Sizov, G., Gambäck, B., Lynum, A.: NTNU-CORE: combining strong features for semantic similarity. In: In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics (2013)
40. Meadow, C.T.: Text Information Retrieval Systems. Academic Press, Inc., Cambridge (1992)
41. Mihalcea, R.: Semcor semantically tagged corpus. Unpublished manuscript (1998)
42. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI, vol. 6, pp. 775–780 (2006)

43. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
44. Mitchell, J., Lapata, M.: Composition in distributional models of semantics. Cogn. Sci. **34**(8), 1388–1429 (2010)
45. Moschitti, A.: Efficient convolution kernels for dependency and constituent syntactic trees. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) ECML 2006. LNCS (LNAI), vol. 4212, pp. 318–329. Springer, Heidelberg (2006). https://doi.org/10.1007/11871842_32
46. Niles, I., Pease, A.: Towards a standard upper ontology. In: Proceedings of the International Conference on Formal Ontology in Information Systems-Volume 2001, pp. 2–9. ACM (2001)
47. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
48. Pilehvar, M.T., Jurgens, D., Navigli, R.: Align, disambiguate and walk: a unified approach for measuring semantic similarity. In: ACL, vol. 1, pp. 1341–1351 (2013)
49. Plotkin, G.D.: A note on inductive generalization. Mach. Intell. **5**(1), 153–163 (1970)
50. Popescu, O.: Learning corpus pattern with finite state automata. In: Proceedings of the ICSC 2013 (2013)
51. Popescu, O., Cabrio, E., Magnini, B.: Textual entailment using chain clarifying relationships. In: Proceedings of the IJCAI Workshop Learning by Reasoning and its Applications in Intelligent Question-Answering (2011)
52. Popescu, O., Cabrio, E., Magnini, B.: Textual entailment using chain clarifying relationships. In: Proceedings of FAM-LbR/KRAQ'11, ijcai-11 (2011)
53. Popescu, O., Magnini, B.: Sense discriminative patterns for word sense disambiguation. In: SCAR Workshop, NODALIDA (2007)
54. Popescu, O., Palmer, M., Hacks, P.: Mapping CPA onto ontonotes. In: Proceedings of the 9th International Conference on Language Resources and Evaluation - LREC14 (to appear)
55. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007 (1995)
56. Sahlgren, M.: The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces (2006)
57. Salton, G., McGill, M.J.: Introduction to modern information retrieval (1983)
58. Šarić, F., Glavaš, G., Karan, M., Šnajder, J., Bašić, B.D.: Takelab: systems for measuring semantic text similarity. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 441–448. Association for Computational Linguistics (2012)
59. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, vol. 12, pp. 44–49 (1994)
60. Shareghi, E., Bergler, S.: CLaC-CORE: exhaustive feature combination for measuring textual similarity. In: In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics (2013)
61. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231 (2006)

62. Sultan, M.A., Bethard, S., Sumner, T.: Back to basics for monolingual alignment: exploiting word similarity and contextual evidence. Trans. Assoc. Comput. Linguist. **2**, 219–230 (2014)
63. Sultan, M.A., Bethard, S., Sumner, T.: Dls@cu: Sentence similarity from word alignment. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), p. 241 (2014)
64. Sultan, M.A., Bethard, S., Sumner, T.: Dls@cu: sentence similarity from word alignment and semantic vector composition. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 148–153 (2015)
65. Surdeanu, M., Ciaramita, M., Zaragoza, H.: Learning to rank answers to non-factoid questions from web collections. Comput. Linguist. **37**(2), 351–383 (2011)
66. Turney, P.D., Pantel, P.: From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. **37**(1), 141–188 (2010)
67. Vo, N.P.A., Caselli, T., Popescu, O.: FBK-TR: applying SVM with multiple linguistic features for cross-level semantic similarity. In: SemEval 2014, p. 284 (2014)
68. Vo, N.P.A., Popescu, O.: Corpora for learning the mutual relationship between semantic relatedness and textual entailment. In: The 10th International Conference on Language Resources and Evaluation (LREC) (2016)
69. Vo, N.P.A., Popescu, O.: A multi-layer system for semantic textual similarity. In: The 9th International Joint Conference on Knowledge Discovery and Information Retrieval (KDIR) (2016)
70. Vo, N.P.A., Popescu, O., Caselli, T.: FBK-TR: SVM for semantic relatedness and corpus patterns for RTE. In: SemEval 2014, p. 289 (2014)
71. Weischedel, R., et al.: Ontonotes: a large training corpus for enhanced processing. In: Handbook of Natural Language Processing and Machine Translation. Springer, Heidelberg (2011)
72. Wise, M.J.: String similarity via greedy string tiling and running Karp-Rabin matching. Online Preprint, Dec 119 (1993)
73. Wise, M.J.: Yap 3: improved detection of similarities in computer program and other texts. In: ACM SIGCSE Bulletin, vol. 28, pp. 130–134. ACM (1996)
74. 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 (1994)
75. Zanzotto, F.M., Dell'Arciprete, L.: Distributed tree kernels. arXiv preprint arXiv:1206.4607 (2012)

Cell Classification for Layout Recognition in Spreadsheets

Elvis Koci[1,2(✉)], Maik Thiele[1], Oscar Romero[2], and Wolfgang Lehner[1]

[1] Database Technology Group, Department of Computer Science,
Technische Universität Dresden, Dresden, Germany
{elvis.koci,maik.thiele,wolfgang.lehner}@tu-dresden.de
[2] Departament d'Enginyeria de Serveis i Sistemes d'Informaciò,
Universitat Politecnica de Catalunya, Barcelona, Spain
{ekoci,oromero}@essi.upc.edu

Abstract. Spreadsheets compose a notably large and valuable dataset of documents within the enterprise settings and on the Web. Although spreadsheets are intuitive to use and equipped with powerful functionalities, extracting and reusing data from them remains a cumbersome and mostly manual task. Their greatest strength, the large degree of freedom they provide to the user, is at the same time also their greatest weakness, since data can be arbitrarily structured. Therefore, in this paper we propose a supervised learning approach for layout recognition in spreadsheets. We work on the cell level, aiming at predicting their correct layout role, out of five predefined alternatives. For this task we have considered a large number of features not covered before by related work. Moreover, we gather a considerably large dataset of annotated cells, from spreadsheets exhibiting variability in format and content. Our experiments, with five different classification algorithms, show that we can predict cell layout roles with high accuracy. Subsequently, in this paper we focus on revising the classification results, with the aim of repairing misclassifications. We propose a sophisticated approach, composed of three steps, which effectively corrects a reasonable number of inaccurate predictions.

Keywords: Speadsheet · Tabular · Table · Document · Layout Recognition · Analysis · Classification

1 Introduction

Spreadsheet applications have evolved to be a tool of great importance for transforming, analyzing, and representing data in visual way. In industry, a considerable amount of the enterprise knowledge is stored and managed in this format. Domain experts use spreadsheets for financial analysis, logistics and planning. Also, spreadsheets are a popular format on the Web. Of particular importance are those that can be found in Open Data platforms, where governments, important institutions, and non profit organizations are making their data available.

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 78–100, 2019.
https://doi.org/10.1007/978-3-319-99701-8_4

All this make spreadsheets a valuable source of information. However, they are optimized to be user-friendly rather than machine-friendly. The same data can be formatted in different ways depending on the information the user wants to convey. It is relatively easy for humans to interpret the presented information, but it is rather hard to do the same algorithmically. As a result, we are constrained to cumbersome approaches that limit the potential reuses of data maintained in these files. A typical problem that arises in most enterprises is that due to the lack of visibility the data stored in spreadsheets is not available for enterprise-wide data analysis or reuse.

Our goal is to overcome these limitations by developing a method that allows to discover tables in spreadsheets, infer their layout and other implicit information. We believe that this approach can provide the means to extract a richer and more structured representation of data from spreadsheets. This representation will act as the base for transforming the data into other formats, such as a relational table/s or a JSON documents.

In this paper we discuss an extended version of our work, which was first introduced at [1]. Here, we focus as well on layout inference via cell classification, describing all the important aspect of our approach. However, we also put emphasis on the post-classification process, where we attempt to correct incorrect predictions.

The paper is organized as follows: In Sect. 2, we define the classification problem for layout inference in spreadsheets. We present the dataset used as ground truth, in Sect. 3. We describe, in Sect. 4, our cell classification approach. Here, we list all defined cell features, explain how the most promising were selected, and provide the evaluation results from the classification experiments. A thorough report of our misclassification repairing approach can be found Sect. 5. Finally, we review related work on table identification and layout discovery in Sect. 6.

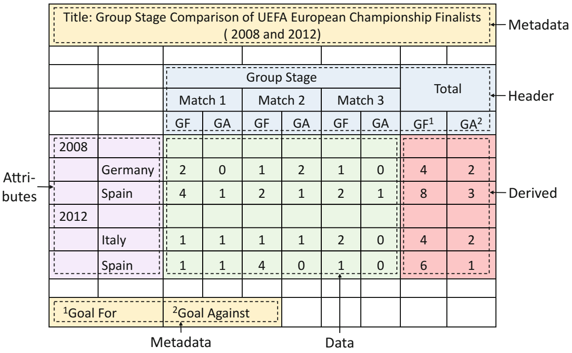
2 The Classification Problem

The objective of capturing the tabular data embedded in spreadsheets can be treated as a classification problem where the individual sections of a table have to be identified. In this section, initially, we define these sections or building blocks that form our classes. Subsequently, we specify the granularity on which the classification task will be performed.

2.1 Spreadsheet Layout Building Blocks

Considering that tables embedded in spreadsheets vary in shape and layout, it is rather challenging to directly recognize them as a whole. Thus, we opt to recognize their building blocks, instead.

We define five building blocks for spreadsheet tables: Headers, Attributes, Metadata, Data and Derived (see Fig. 1). A “Header” (H) cell represents the label of a column and can be flat or hierarchical (stacked). Hierarchical structures



Title: Group Stage Comparison of UEFA European Championship Finalists (2008 and 2012)									
		Group Stage						Total	
		Match 1		Match 2		Match 3			
		GF	GA	GF	GA	GF	GA	GF[1]	GA[2]
2008									
	Germany	2	0	1	2	1	0	4	2
	Spain	4	1	2	1	2	1	8	3
2012									
	Italy	1	1	1	1	2	0	4	2
	Spain	1	1	4	0	1	0	6	1
[1]Goal For		[2]Goal Against							

Fig. 1. The building blocks [1].

can be also found in the left-most or right-most columns of a table, which we call "Attributes" (A), a term first introduced at [2]. Attributes can be seen as instances from the same or different (relational) dimensions placed in one or multiple columns in a way that conveys the existence of a hierarchy. We label cells as "Metadata" (M) when they provide additional information regarding the worksheet as a whole or specific sections. Examples of Metadata are the table name, creation date, and the unit of the values for a column. The remaining cells form the actual payload of the table and are labeled as "Data". Additionally, we use the label "Derived" (B) to distinguish those cells that are aggregations of other Data cells' values. Derived cells can have a different structure from the core Data cells, therefore we need to treat them separately. Figure 1 provides examples of all the aforementioned building blocks.

2.2 Working at the Cell Granularity

One potential solution for the table identification and layout recognition tasks is to operate under some assumptions about the structure of spreadsheet tables. That means expecting spreadsheets to contain one or more tables with typical layouts that are well separated from each other. In such scenario we could define simple rules and heuristics to recognize the different parts. For example, the top row could be marked as Header when it contains mostly string values. Additionally, cells containing the string "Table:" are most probably Metadata. However, this approach can not scale to handle arbitrary spreadsheet tables. Since, the corpora we have considered include spreadsheets from various domains, we need to find a more accurate and more general solution.

For this reason, our approach focuses on the smallest structural unit of a spreadsheet, namely the cell. At this granularity we are able to identify arbitrary layout structures, which might be neglected otherwise. For instance, it is tricky to classify rows when multiple tables are stacked horizontally. The same applies for the cases when Metadata are intermingled with Header or Data. Nevertheless, we acknowledge that the probability of having misclassifications

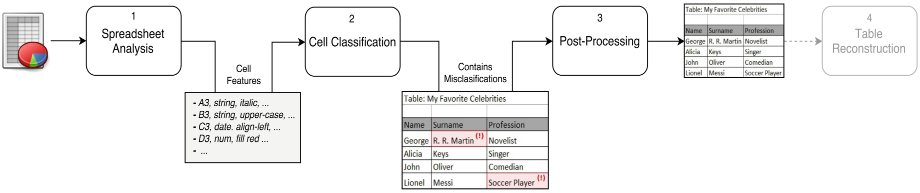

1
Spreadsheet Analysis
Cell Features
- A3, string, italic, ...
- B3, string, upper-case, ...
- C3, date. align-left, ...
- D3, num, fill red ...
- ...
2
Cell Classification
Contains Misclasifications
Table: My Favorite Celebrities
Name | Surname | Profession
George | R. R. Martin (!) | Novelist
Alicia | Keys | Singer
John | Oliver | Comedian
Lionel | Messi | Soccer Player (!)
3
Post-Processing
Table: My Favorite Celebrities
Name | Surname | Profession
George | R. R. Martin | Novelist
Alicia | Keys | Singer
John | Oliver | Comedian
Lionel | Messi | Soccer Player
4
Table Reconstruction

Fig. 2. The cell classification process [1].

increases when working with cells instead of composite structures such as rows or columns. Therefore, our aim is to come up with novel solutions that mitigate this drawback.

Figure 2 illustrates the three high-level tasks that compose our cell classification process. Initially, the application reads the spreadsheet file and extracts the features of each non-blank cell. Here we considered different aspects of the cell, summarized in Sects. 4.1 and 4.2. In the next step, cells are classified with high accuracy (see Sect. 4.3) based on their features. Finally, a post-processing step improves even further the classification results. Using rules and machine learning techniques we identify cells that are most probably misclassified, and attempt to infer their true label (i.e., layout role).

To complete the picture, Fig. 2 also includes the Table Reconstruction task, which forms a separate topic, and is therefore left as future work.

3 The Gold Standard

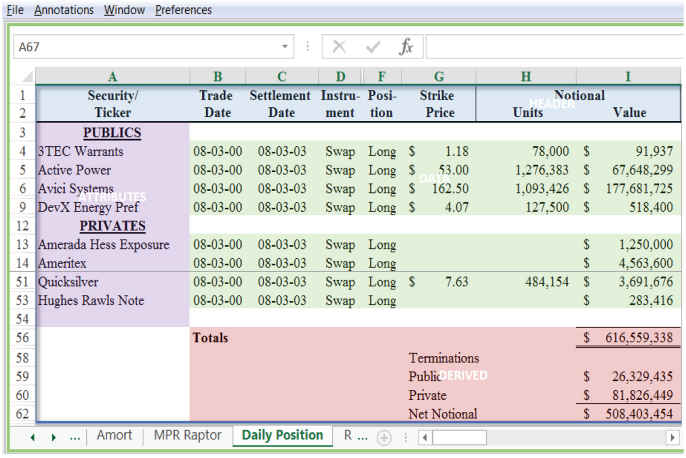
The supervised classification processes requires a ground truth dataset, which is used for both training and validation. In Sect. 3.1 we briefly describe the three spreadsheet corpora used to extract a representative set of spreadsheets. To create the training data we developed a spreadsheet labeling tool (see Sect. 3.2), which provides the means to annotate ranges of cells. Given that tool, we randomly selected and annotated files from the three corpora for which we provide statistics in Sect. 3.3.

3.1 Spreadsheet Corpora and Training Data

For our experiments we have considered spreadsheets from three different sources. EUSES [3] is one of the oldest and most frequently used corpora. It has 4, 498 unique spreadsheets, which are gathered through Google search, using keywords such as “financial” and “inventory”. The ENRON corpus [4] contains over 15, 000 spreadsheets, extracted from the Enron email archive. This corpus is of a particular interest, since it provides access to real-world enterprise spreadsheets. The third corpus is FUSE [5] that contains 249, 376 unique spreadsheets, extracted from Common Crawl. Each spreadsheet in FUSE is accompanied by a JSON file that contains metadata and statistics. Unlike the other two corpora, FUSE can be reproduced and extended.

3.2 The Annotation Tool

Using the Eclipse SWT[1] library we developed an interactive desktop application that ensures good quality annotations. The original Excel spreadsheet is embedded into a Java window and protected from user alteration. To create an annotation, the user selects a range of cells (region) and then chooses the appropriate predefined label. A rectangle, which is filled with the color associated to the label, covers the annotated region. The application evaluates the annotations and rejects the inappropriate ones. For example, the user cannot annotate ranges that are empty or overlapping with existing annotations. The data from all the created annotations are stored in a new sheet, named "Range_Annotation_Data". The sheet is protected and hidden once the file is closed. Figure 3 provides an example of an annotated sheet.



	A	B	C	D	F	G	H	I
1	Security/	Trade	Settlement	Instru-	Posi-	Strike	Notional	
2	Ticker	Date	Date	ment	tion	Price	Units	Value
3	PUBLICS							
4	3TEC Warrants	08-03-00	08-03-03	Swap	Long	$ 1.18	78,000	$ 91,937
5	Active Power	08-03-00	08-03-03	Swap	Long	$ 53.00	1,276,383	$ 67,648,299
6	Avici Systems	08-03-00	08-03-03	Swap	Long	$ 162.50	1,093,426	$ 177,681,725
9	DevX Energy Pref	08-03-00	08-03-03	Swap	Long	$ 4.07	127,500	$ 518,400
12	PRIVATES							
13	Amerada Hess Exposure	08-03-00	08-03-03	Swap	Long			$ 1,250,000
14	Ameritex	08-03-00	08-03-03	Swap	Long			$ 4,563,600
51	Quicksilver	08-03-00	08-03-03	Swap	Long	$ 7.63	484,154	$ 3,691,676
53	Hughes Rawls Note	08-03-00	08-03-03	Swap	Long			$ 283,416
54								
56		Totals						$ 616,559,338
58						Terminations		
59						Public		$ 26,329,435
60						Private		$ 81,826,449
62						Net Notional		$ 508,403,454

Fig. 3. Annotated sheet [1]. (Color figure online)

In addition to the building block described in Sect. 2, we have introduced the possibility to annotate a region (area) that represents a whole table. A rectangle with thick blue borders marks its boundaries.

We need the "Table" annotation for two main reasons: Firstly, we can govern the labeling process, to assure valid annotations. For example, Data can only exist inside a Table. However, Metadata can be left outside when they provide information relevant to multiple Tables. Secondly, these annotations will help us evaluate the Table Reconstruction task, in the future.

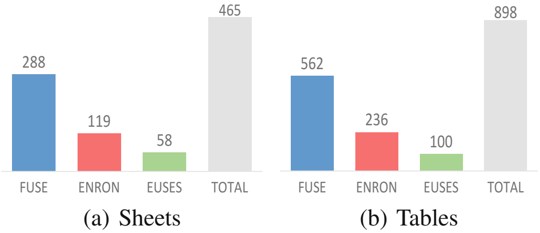
[1] https://www.eclipse.org/swt/.

3.3 Annotation Statistics

The graphs below provide an overview of the collected annotations and the contribution of each corpus. We considered each corpus individually and assigned a unique number to their files. Using a random number generator we extracted subsets of files. From these, we annotated a total of 465 worksheets (216 files) and 898 tables (Fig. 4).

In Fig. 5 we examine the annotated cells. The total number of cells for each label (class) is placed at the top of the column bar. There was a small amount of cells that did not match any of the defined labels. These are usually random notes that do not have a clear role, and do not provide additional context (information) about the table. We decided to omit such cells.

As can be seen in Fig. 5, the number of Data cells is by orders of magnitude larger than the other label numbers. To adjust the class distribution we undersampled the Data class. We consider from each table in our dataset the first, the last row, and three random rows in between. By applying this technique the Data class was reduced to 32, 875 instead of 808, 179 cells. Considering also the other four classes the final gold standard consists of 52, 948 cells in total.



(a) Sheets (b) Tables

Fig. 4. Annotation statistics [1].

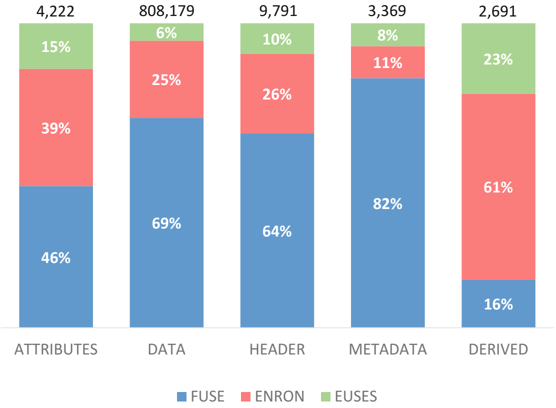


Fig. 5. Annotated cells [1].

4 Cell Classification

4.1 Feature Specification

We have grouped the defined features into 5 categories: content, cell style, font style, reference, and spatial. The *content* features describe the cell value, but not its format. The *cell style* features capture the formatting applied to the cell, excluding the font formatting. The later is recorded by *font style* features. *Reference* features explore the Excel formulas and their references in the same or other worksheets. The last group, *spatial* features, describe the "neighborhood" of the cell (i.e., the adjacent cells). Here, we do not define the individual features, instead more details can be found at [1].

4.2 Feature Selection

We used Weka[2], a well known tool for machine learning tasks, for feature selection and classification. Initially, we binarized nominal features with more than two values, which gave us 219 features in total. We used the "RemoveUseless" option to remove the features that do not vary at all or vary too much. Additionally, we manually removed those features that are practically constant (i.e., at least 99.9% of cases the value is the same). Furthermore, we decided to exclude from the final set features that check the style and content type of the neighbors. Although, these features are promising, they need further refinement. Thus, we plan perform thorough experiments with them in the future.

The remaining 88 features were evaluated using the *InfoGainAttribute*, *GainRatioAttribute*, *ChiSquaredAttribute*, *ConsitencySubset*, and *CfsSubset* feature selection methods. For each one of them we performed 10 folds (runs). A bidirectional Best First search was used for *ConsitencySubset* and *CfsSubset*, while the other methods can only be coupled with Ranker search.

From the results we considered features that score high in all these selection methods. Although, we were predominantly influenced by *ConsistencySubset* results, since, when tested, they provide the highest classification accuracy. We also included in the final set features that are strong indicators despite the fact that they describe small number of instances. "Words_Like_Table" is an example of such features, where 48 out of total 49 positive (true) cases are instances of the Metadata class.

Tables 1 and 2 list the selected features, 43 in total. Those suffixed with ? represent boolean features. While, those suffixed with # represent numeric features.

In general spreadsheets exhibit different characteristics depending on the domain they come from. Therefore, we expect that some of this features might not works as well for other spreadsheet datasets. For example, reference features are more important for industrial rather than for Web spreadsheets, since the former are characterized by heavier use of formulas. We note that an independent feature selection might be required for other spreadsheet datasets, in order to achieve near optimum accuracy.

[2] http://www.cs.waikato.ac.nz/ml/weka/.

Table 1. Selected content and style features [1].

Content	Cell style
LENGTH#	INDENTATIONS#
NUMBER_OF_TOKENS#	H_ALIGNMENT_DEFAULT?
LEADING_SPACES#	H_ALIGNMENT_CENTER?
IS_NUMERIC?	V_ALIGNMENT_BOTTOM?
IS_FORMULA?	FILL_PATTERN_DEFAULT?
STARTS_WITH_NUMBER?	IS_WRAPTEXT?
STARTS_WITH_SPECIAL?	NUMBER_OF_CELLS#
IS_CAPITALIZED?	NONE_TOP_BORDER?
IS_UPPER_CASE?	THIN_TOP_BORDER?
IS_ALPHABETIC?	NONE_BOTTOM_BORDER?
CONTAINS_SPECIAL_CHARS?	NONE_LEFT_BORDER?
CONTAINS_PUNCTUATIONS?	NONE_RIGHT_BORDER?
CONTAINS_COLON?	MEDIUM_RIGHT_BORDER?
WORDS_LIKE_TOTAL?	HAS_NO_DEFINED_BORDERS?
WORDS_LIKE_TABLE?	
IN_YEAR_RANGE?	

Table 2. Selected font, reference and spatial features [1].

Font	Reference	Spatial
FONT_SIZE#	IS_AGGREGATION_FORMULA?	ROW_NUMBER#
FONT_COLOR_DEFAULT?	REFERENCE_VALUE_NUMERIC?	COLUMN_NUMBER#
IS_BOLD?		HAS_0_NEIGHBORS?
NONE_UNDERLINE?		HAS_1_NEIGHBOR?
		HAS_2_NEIGHBORS?
		HAS_3_NEIGHBORS?
		HAS_4_NEIGHBORS?

4.3 Cell Classifiers

In our evaluation, we considered various classification algorithms, most of which have been successfully applied to similar tasks in the literature. Specifically, we considered CART [6] (*SimpleCART* in Weka), C4.5 [7] (*J48* in Weka), Random Forest [8] and support vector machines [9] (*SMO* in Weka). The latter uses the sequential minimal optimization algorithm developed by [10] to train the classifier. Here, with SMO we considered both polynomial kernel and RBF kernel.

We evaluated the classification performance using 10-fold cross validation. The Random Forest (RF) gave the highest overall accuracy of 98.2%. Also, RF outperformed the other algorithms, in all the defined classes. To provide a more concrete picture on the classification accuracy, we also tested Random Forest against the full dataset of annotated cells (828, 252). There was a slight decrease in performance specifically for Metadata and Derived, but the general accuracy did not suffer. More detailed results can be found at [1].

5 Post-processing

In this section, we discuss techniques for handling the misclassifications that occur during the cell classification process. It is in our benefit to revise these incorrect predictions as it will make the subsequent tasks, such as table identification and schema extraction in spreadsheets, much more easier and accurate. In the following paragraphs, we discuss two approaches. We start with what can be considered a naïve approach, and afterwards we motivate and describe our more sophisticated solution.

5.1 Naïve Approach

Our initial empirical analysis of the classification results hinted that neighboring cells could be potentially used to recover some of the misclassifications. Intuitively, the label assigned to a cell should match, in most of the scenarios, at least that of the neighboring cells in the same row and/or column.

Here, we define the neighborhood as a 3-by-3 window around a cell, shown in Fig. 6. The red cell in the center, marked with "x", represents a misclassification, surrounded by 8 neighboring cells.

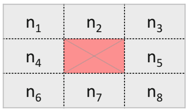


Fig. 6. A 3 × 3 cell neighborhood. (Color figure online)

Not necessarily all neighboring cells have a label. For example, empty and hidden cells do not get labeled, since we omit them from the classification process. Also, another special case are the misclassified cells at the boundaries (extremes) of the worksheet (i.e., the minimum and maximum allowed row/column in the spreadsheet application). Such cells have less than eight neighbors, since one of the neighboring columns and/or rows does not exist.

We would like to standardize the number of neighbors for any arbitrary cell to eight. We accomplish this by adding two artificial labels: "Empty" and "Imaginary". The latter shall be used for (non-existing) neighbors outside the boundaries of the worksheet, while the first for all the other cases of un-labeled neighboring cells.

Having seven labels in total, we implemented a script to find distinct arrangements of labels in the neighborhood of incorrectly classified cells. From 1, 237 misclassifications, resulting from the classifications in the full dataset (828, 252 cells), there were 672 unique neighborhood label arrangements.

We identified the first 40 most repeated arrangements and manually inferred generic rules (i.e., not bound to specific labels) from them. These rules are divided into two sets: identification and relabeling. Intuitively, we use the first rule-set to identify incorrect classifications, and afterwards we relabel them using the second rule-set. Using this technique we managed to repair 152 misclassified, but lose 14 correctly classified cells. Here, we do not provide details about the individual rules, instead the complete list can be found at [1].

Though this method is able to recover a number of incorrect classifications, it has several limitations. Our subsequent experiments revealed that in almost half of the cases there are misclassified cells in the neighborhood itself, as illustrated in Fig. 7. Beside these observations, there are other good reasons to look further than the immediate neighborhood. For example, a Header cell that is relatively far from all Data cells in a worksheet is probably a misclassification. Another example comes from tables with missing values, which translates at Data cells having empty immediate neighbors. In such scenario, we need to look at more distant neighbors to determine if a Data cell is misclassified or not.

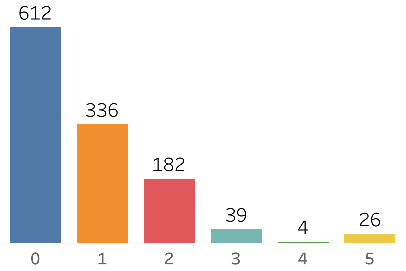
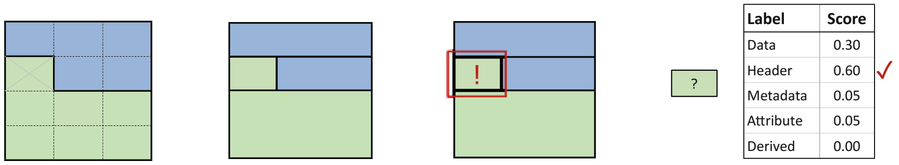


Fig. 7. Occurrences of misclassifications in the immediate neighborhood.

In the following section we propose a new approach for detecting and fixing misclassified cells. This approach is based on a good mixture of features, extracted from the immediate and the distant neighbors. The results from our evaluation are very encouraging and confirm the advantages of taking into account a broader neighborhood context.

5.2 Region Based Approach

In this section we present our region-based approach (RBA) for handling misclassifications. At the core of RBA, is the intuition that grouping adjacent cells of the same label should form regions of rectangular shape. That is because tables have well separated sections, such as the Data and Header sections, which tend to be rectangular instead of an arbitrary rectilinear polygonal shape. Here, we consider as adjacent cells only immediate neighbors on the left, right, top, and bottom (see Fig. 6). Additionally, we define rectangular region as a well formed matrix of cells of the same label. Intuitively, in such a matrix, all rows have the same number of cells. The same is true when we examine the columns of the matrix.



Label	Score
Data	0.30
Header	0.60
Metadata	0.05
Attribute	0.05
Derived	0.00

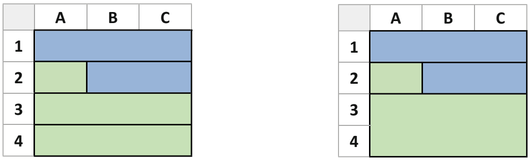
(a) Load (b) Standardize (c) Identify (d) Relabel

Fig. 8. Region based approach. (Color figure online)

We claim that non-rectangular cell regions point towards misclassifications. In other words, some cell/s break the regularity of the region. For us these cells are potential misclassifications. Hence, the aim of our approach becomes to isolate them, and subsequently determine the correctness of their label. Figure 8a illustrates how a misclassification, the cell marked with an x, can introduce irregularities.

Also, in Fig. 8 we introduce the tasks that compose RBA. Initially, we load the classification results. We then create strictly rectangular regions per each label. In the subsequent step, we identify incorrect classifications. Finally, we predict the true label for those regions flagged as misclassified.

Standardization and Confinement. This task starts by building what we call Row Label Intervals (RLI). We define these intervals as a sequence of cells of the same label in a row. Figure 9a displays the intervals from the example shown in Fig. 8a. The first, third, and fourth row contain one interval each, while the second row contains two intervals of different labels.



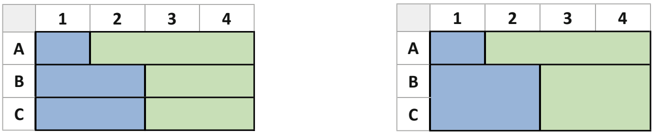
(a) Row Intervals (b) Regions

Fig. 9. Original worksheet. (Color figure online)

As we emphasized in the previous section, we are interested in strictly rectangular (cell) regions. We try to achieve this by merging RLIs of same label. This is only possible when the RLIs in adjacent rows have the same start and end. Otherwise we introduce shape irregularities. Based on this reasoning, for our running example, we merge the intervals in the third and fourth row, as illustrated in Fig. 9b. Using this technique we manage to isolate the single-cell region that prevented the rest of the green cells to form a well-shaped cluster.

We were not able to merge the blue row intervals from 1st and 2nd row, since they have a different start column. However, clearly there is the potential to build a larger blue region, that is B1:C2. This would have isolated the cell A1. The latter is desirable, since at this phase we aim at pinpointing irregularities.

One way to tackle this challenge is by creating regions column-wise, in addition to row-wise. We pivot (transpose) the worksheet, so that the columns of the original worksheet become the rows of the transformed worksheet. Afterwards, it is easy to construct row intervals, following the steps described previously. The results are shown in Fig. 10a. Once we attempt to merge the RLIs, we get the output shown in Fig. 10b. As intended we create a large blue region, and isolate the blue interval that does not fit well with the rest.



(a) Pivot Intervals

(b) Regions

Fig. 10. Pivoted worksheet. (Color figure online)

Our standardization procedure produces two alternative partitioning strategies for the labeled cells. For some worksheets the optimum partitions might come from one of the directions (i.e., row-wise or column-wise). However, for others we need both directions to ensure that for each misclassified cell there is at least a region that confines it from correct classifications. Thus, we have to keep both outputs, which has the drawback of augmenting the number of regions in the worksheet. One possible and required action for reducing this number is to filter out duplicate regions from the outputs. Figure 11 summarizes the overall standardization procedure, which includes the duplicates filtering step.

Confinement Assessment. To assess the validity of this procedure we decided to evaluate it on the classification results. We divide the resulting regions into three categories based on the ratio of misclassified cells they contain. We call “Correct” those regions that do not contain any misclassification. For those that only contain misclassifications we use the term “Misclassified”. The remaining cases, regions that contain both correct and incorrect classifications, we call “Mixed”. Figure 12a provides the number of regions per category.

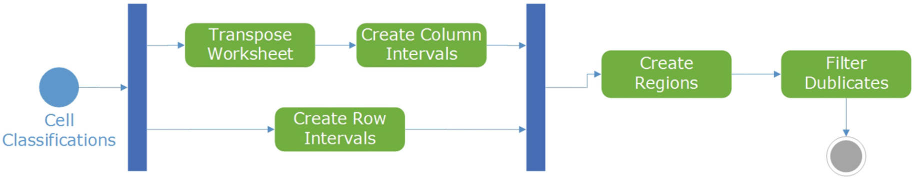
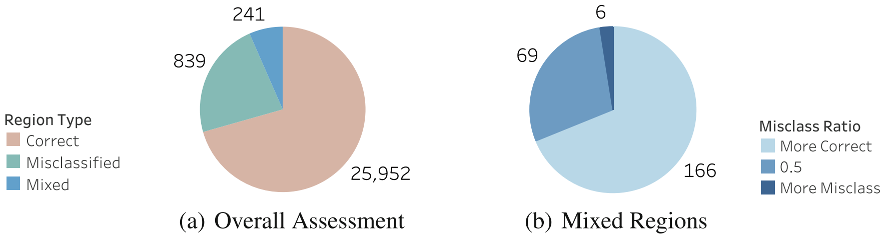


Fig. 11. Standardization procedure.

Mixed regions might raise concerns at first glance, since they occur a noticeable number of times. However, they are a natural by-product of the procedure. Consider again Fig. 10b, the first cell of the green region (interval) at the top row is misclassified, as pointed out in Fig. 8a. In this example, the Mixed region occurs when processing the transformed (pivoted) worksheet. However, in the general case both variants of the worksheet can produce such regions.

Additionally, we have performed a more detailed analysis of Mixed regions. The results are displayed in Fig. 12b. We note that for the majority of cases the number of correctly classified cells is greater than that of misclassified cells. Also, there are 69 cases where the numbers are equal, and an insignificant number of cases with more misclassified cells.



(a) Overall Assessment (b) Mixed Regions

Fig. 12. Region analysis.

To simplify our subsequent operations, we decided to maintain only two categories of regions. Those that mostly contain incorrect classifications (>0.5) are marked as Misclassified, the rest as Correct. These brings the number of regions per category to 845 and 26, 187 respectively.

Filtering by Size. As mentioned before, one of the drawbacks of our standardization procedure is the considerable number of outputted regions. Ideally, we would like to keep only those that have the most potential to be Misclassified. Therefore, we analyzed the Misclassified regions, with the purpose of identifying some of their typical characteristics. Our analysis revealed that these regions exhibit small sizes (i.e., number of cells in the region), as shown in Fig. 13. Clearly, the larger is the size of a Misclassified region the least are the occurrences.

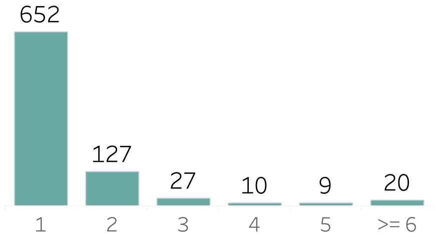


Fig. 13. Size occurrences in misclassified regions.

Based on the results in Fig. 13, we decided to consider only regions of size 1–3 from the procedure outputs, since they have significant number of occurrences. After, performing this filter, for our dataset, we get $12,724$ regions. Out of these, 806 are Misclassified, and $11,918$ are Correct regions. We utilize this reduced dataset for the tasks described in the following sections.

Detecting and Repairing Misclassifications. We have defined misclassification detection and repairing (relabeling) as supervised learning tasks. For this, we have created a set of features, which are formally described in the following paragraphs. Most of these features are used for both detection and repairing.

Region Features. Table 3 summarizes the features that are used to characterize each rectangular region. We use the same convention as in Sect. 4.2 to distinguish numerical features from boolean ones. We note that *predicted_label* does not fit in any of these categories, since it is a nominal feature. Here, we have additionally introduced the categories “Simple” and “Compound”. As their formal definition will show, compound features are derived from multiple simple ones (some of them not explicitly listed in the table).

All features, introduced in this section, are based on the conception that a region can be solely represented with the rectangle that bounds its cells and the label of these cells. The worksheet itself can be seen as a Cartesian Coordinate system, where the point $(1, 1)$ is at the top left corner. The values of the x-axis increase column-wise, while for y-axis they increase row-wise.

Having such coordinate system, it is relatively easy to convert the regions into abstract rectangles. The coordinates for the top-left vertex of the rectangle are the column and row number of the top-left cell in the region. Its width and height can be calculated by counting respectively the number of cells in a column and in a row of the region. For example, the large green region in Fig. 9b will be represented with the rectangle having as top-left vertex the point $(1, 3), width = 3$, and $height = 2$.

The *simple features* characterize various aspects of the rectangle (region). A rectangle is horizontal when $width > height$, is vertical when $width < height$, and is square when $width = height$. The feature *count_cell* describes how many cells are in the region (i.e., the area of the rectangle). Additionally, we count the

Table 3. Region features.

Nr.	Simple	Nr.	Compound
1	IS_HORIZONTAL?	9	SIMILARITY_{TOP,BOTTOM,LEFT,RIGHT}#
2	IS_VERTICAL?	10	DISSIMILARITY_{TOP,BOTTOM,LEFT,RIGHT}#
3	IS_SQUARE?	11	EMPTINESS_{TOP,BOTTOM,LEFT,RIGHT}#
4	COUNT_CELLS#	12	INFLUENCE_{TOP,BOTTOM,LEFT,RIGHT}#
5	COUNT_ITS_KIND#		
6	DISTANCE_FROM_ITS_KIND#		
7	DISTANCE_FROM_ANY_KIND#		
8	PREDICTED_LABEL		

number of regions in the worksheet having the same label (i.e., its own kind) as the current region. Simple feature number 6 and 7 respectively capture the smallest Euclidean distance of this region to a region of the same label and to a region of any label. Finally, the *predicted_label* stores the label (i.e., assigned from the cell classification task as described in Sect. 4) of the cells in the region.

With the *compound features* we analyze the neighborhood of a region, similarly to the naïve approach (see Sect. 5.1). However, this time the neighborhood is made of other regions, instead of cells. Therefore, below we refine some of the previously used concepts and add some new ones.

- **Current Region (R):** The region whose neighborhood we are studying.
- **Direction (D):** Can be Top, Bottom, Left, or Right.
- **Neighbor (N):** Any region other than the current one.
- **Nearest neighbors (NNs):** The neighboring regions with the smallest Euclidean distance from the current region in the specified direction.
- **Similar neighbors (SNs):** Neighbors that have the same label as the current region.
- **Dissimilar neighbors (DNs):** Neighbors that have different label from the current region.

As shown in the above list, we study the neighborhood in four directions, omitting intermediate directions like top-right. Moreover, we use the concept of nearest neighbor, to distinguish from other neighbors in the vicinity of the current region. However, as we shall see in the following paragraphs, we consider more distant neighbors as well. Additionally, we examine the label of the neighbors, to determine if they are similar or dissimilar to the current region. Further more, we are interested in regions with specific label, as we later show in the definition of *influence*.

It is important to emphasize that unlike the naïve approach, the number of neighboring regions varies. Moreover, they might come in different sizes (considering both width and height). Therefore, we need a method to weight the importance of each neighbor. For this we utilize two measures: *overlap-ratio* and *distance*. The former quantifies how much of the specified direction is dominated by a neighbor. The latter how far or close is the neighbor. Clearly, a nearer neighbor should be given more weight.

Equation 1, illustrates how the overlap ratio is calculated. Here, r stands for the current rectangle, n for the selected neighbor, and d for the current direction. For a neighbor at the top and bottom we measure the overlap by projecting it and the region itself to the x-axis. The length of the segment shared by both these regions represents the overlap. For left and right neighbors we do the same, but instead using the projections to the y-axis. We transform the overlap into a ratio by dividing it with the width or the height (i.e., respectively, the length of the vertical or horizontal edge) of the current region.

$$\begin{gathered} OverlapRatio(r, n_i, d) = \frac{Overlap(r, n_i, d)}{EdgeLength(r, d)} \\ \text{where} \quad n_i \in Neighbors(r, d) \quad \text{and} \quad d \in Directions \end{gathered} \tag{1}$$

Once we have the overlap ratio and the distance to the neighbor, we can calculate its weight as shown in Eq. 2. In the denominator, we add one to the distance to account for cases where the latter is zero. Clearly, this equation captures the intuition that the weight for a neighbor should increase for smaller distances and bigger overlap ratios.

$$weight_i = OverlapRatio(r, n_i, d) \cdot \frac{1}{1 + Distance(r, n_i)} \tag{2}$$

We can now define the *similarity* for a region and its neighbor, as shown in Eq. 3. Similarity takes a value greater than zero when the neighbor is a SN and is one of the NNs. In such scenario, the value of the *similarity* equals the weight of the neighbor.

$$similarity_i = \begin{cases} 0 & Label(r) \neq Label(n_i) \ \vee \ n_i \notin Nearest(r, d) \\ weight_i & otherwise \end{cases} \tag{3}$$

$$dissimilarity_i = \begin{cases} 0 & Label(r) = Label(n_i) \ \vee \ n_i \notin Nearest(r, d) \\ weight_i & otherwise \end{cases} \tag{4}$$

Likewise we calculate the *dissimilarity* for a neighbor, as shown in Eq. 4. The only difference from the definition of *similarity* is that here the neighbor must be a DN, in addition to being a NN.

Influence goes beyond the immediate neighborhood (i.e., the nearest neighbors). It quantifies how much distant neighbors influence the current region. For example, this can be useful in the scenario where two Correct regions of the same label are separated by a Misclassified region. Knowing that there is considerable influence from a SN might save the region from accidentally being marked as Misclassified. Good influence can come also from other labels. For instance, a strong Data influence from the bottom neighborhood, can reinforce the belief that Header is the most plausible label for the current region.

Influence is tightly coupled with the selected label at that time, as shown in Eq. 5. Here, l stands for the label. Also, we have updated the function *Nearest* by adding the optional parameter *label.* When this parameter is set, the function returns only the nearest neighbors for a specific label in the given direction. Influence gets a value greater than zero when there exist at least one neighbor with the requested label. When there are multiple such neighbors, we prefer the influence from the nearest ones.

$$influence_i = \begin{cases} 0 & Label(n_i) \neq l \;\; \vee \;\; n_i \notin Nearest(r, d, l) \\ weight_i & otherwise \end{cases} \tag{5}$$

All the previous equations hint that there can be more than one nearest neighbor. In order to get the total value of a feature for a direction, we need to sum up the values for the individual NNs. Equations 6 and 7 respectively show how to perform this for *similarity* and *influence.* We can calculate the total *dissimilarity* for a direction the same way.

$$total\_similarity_d = \sum_{i=1}^{|Nearest(r,d,Label(r))|} Similarity(r, n_i, d) \tag{6}$$

$$total\_influence_{d,l} = \sum_{i=1}^{|Nearest(r,d,l)|} Influence(r, n_i, d, l) \tag{7}$$

Emptiness, the last compound feature, is the feature that tries to capture the (partial or complete) non-existence of nearest neighbors in a direction. Emptiness takes the maximum value when there are no neighbors in a direction. When the NNs partially overlap with the current region, *emptiness* takes a value between zero and one. Equation 8 illustrates how to calculate the value of this feature for a specific direction. Note in this equation that we do not set the label parameter for the *Nearest* function. Thus, it returns the complete set of NNs.

$$total\_emptiness_d = 1 - \sum_{i=1}^{|Nearest(r,d)|} OverlapRatio(r, n_i, d) \tag{8}$$

We can add additional flavors to the compound features by aggregating them to the level of row (left and right), column (top and bottom), and overall neighborhood (all four directions). Equation 9 illustrates how to calculate the overall value, using as example the *similarity* feature. As shown, we normalize the value from a direction using the ratio of the edge length (in that direction) to the perimeter of the current region.

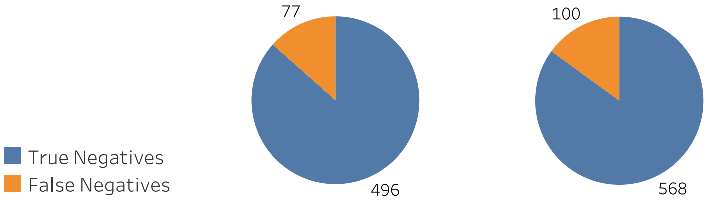
$$overall\_similarity = \sum_{j=1}^{|Directions|} \left(\frac{EdgeLength(r, d_j)}{Perimeter(r)} \cdot similarity_{d_j} \right) \tag{9}$$

Misclassification Identification. We define Misclassification Identification (MI) as a machine learning task, whose goal is to distinguish real Misclassified regions from the dataset of candidate regions. For this binary classification problem we have considered all the simple features mentioned in the previous section. Additionally, we use the compound features for all four directions and their three flavors (i.e., row, column, overall). However, for *influence* we define a special version. We only consider the influence from neighbors of the same label, and omit the influence from the rest. This brings the total number of features, used for MI, to 36 (i.e., 8 simple + 28 compound).

Table 4. Comparing classifiers for misclassification identification task.

	Rand. forest *-I 100*	SMO RBF *-C 19.0, -G 0.1*	Logistic *-R 1.0E-14*	JRIP *-N 10.0*
F1 measure	0.97	0.96	0.95	0.96
True negative rate	0.64	0.58	0.47	0.60
False negative rate	0.03	0.03	0.04	0.04

For our evaluation we experimented with several classification algorithms. Again, we used the Weka tool. We firsed tuned the parameters of the individual classifiers. Subsequently, we used Weka Experimenter[3] to run 10 fold cross-validation with 10 repetitions. The results are displayed in Table 4. The values represent the average of all runs. Random forest achieves the highest F-measure and simultaneously has the highest true negative rate. With what regards false negative rate, there is not substantial difference between the classifiers. Considering these results, we selected the Random Forest classifier for our subsequent analysis.



(a) In Regions (b) In Cells

Fig. 14. Misclassification identification results.

In Fig. 14 we display the results from one of the cross-validation repetitions ($seed = 1$). We have provided the numbers in terms of regions and in terms

[3] https://sourceforge.net/projects/weka/files/documentation/3.8.x/

of individual cells. We get more False Negative cells (i.e., wrongly flagged as Misclassified) in comparison to the naïve approach. However, the number of True Negative cells (i.e., correctly predicted Misclassified) is several times bigger for the RBA approach.

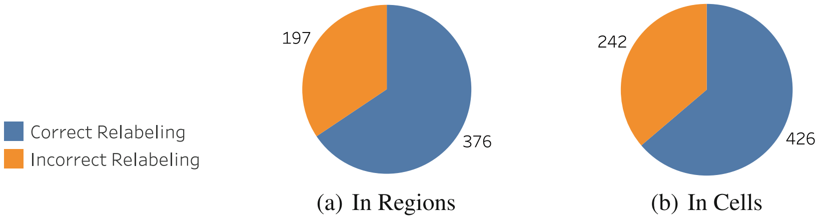
Relabeling. We define the task of relabeling as that of predicting the most plausible true label for regions flagged as Misclassified. For this task we use all the simple features, except of *predicted_label*. From the compound features we use only *influence*, capturing it for each label and direction. We add to the set of directions also the three combined variants (flavors): row, column, and overall. With this addition, the total number of features used for relabeling becomes 42 (i.e., 7 simple + 35 influences).

The dataset to train our model for relabeling comes from the original annotated cells (i.e., the ground truth). Similarly to what we did with the predicted labels, we construct rectangular regions from the annotations. At the end, we keep only those of size three or smaller for training. This brings the total number of regions in this dataset to 11, 934.

Table 5. Relabeling: trained on annotated regions.

	Rand. forest *-I 350*	SMO RBF *-C 16.0 -G 1.0*	Logistic *-R 1.0E-8*	JRIP *-N 2.0*
F1 measure	0.64	0.59	0.67	0.49
True negative rate	0.65	0.59	0.67	0.49
False negative rate	0.11	0.13	0.10	0.16

For our evaluation we used the same classification algorithms as for misclassification identification. In a similar fashion, we first tuned the parameters of the classifiers on the training datasets. Subsequently, we evaluated their performance on the 573 regions identified as Misclassified. The results are provided in Table 5.



(a) In Regions (b) In Cells

Fig. 15. Relabeling results.

Figure 15 displays the relabeling results for Logistic Regression (LR) classifier. We pick this classifier, since as shown in Table 5 it achieves the best results.

Again, we provide the numbers in regions and in cells. Although, we managed to predict the true label for most of the regions flagged Misclassified, there is a considerable number of wrong predictions.

One possible solution to decrease the number of incorrect predictions is to use the predicted class (label) probabilities, instead of fixed membership. By default, the LR classifier assigns the class with the highest probability to an instance. We can interfere in this process, and only relabel those regions for which the prediction has high confidence. Effectively, this means setting a threshold for the class probability.

We assessed the validity of this approach by analyzing the probabilities (scores) assigned by the LR classifier during the relabeling task. For each region we have recorded the highest predicted class score. Then we created the distinct list of these scores from the whole task. For each value in the list we identify the regions that got a score greater than or equal. Then we calculate the *difference* between the number of correctly relabeled and the number of incorrectly relabeled. The results are provided in Fig. 16. The largest *difference* is 192, and is achieved for score 0.59. For this score we get 363 correct versus 171 incorrect predictions. However, we decided to be more conservative and set the (score) threshold 0.83. We get a better trade off, since we only get 113 wrong region (re-)labels, and a considerable number of 300 correct region (re-)labels.

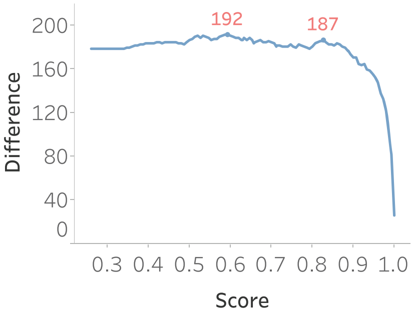


Fig. 16. Confidence score analysis.

6 Related Work

6.1 Spreadsheet Layout Inference

Comparing with related work, we have exclusively focused on the cell level. In this way we can infer the layout of arbitrary tables and arrangements in spreadsheets. Related work proposes approaches that work with larger structures, such as rows and columns. These fail to recognize that the contained cells could have different layout roles. In other words, these approaches lose important information, by assuming homogeneous structures. Furthermore, working on a cell level enables us to make use of many features, not considered before by related work. Currently, there is no scheme that allows to aggregate these features for larger structures without compromising the accuracy. Above all, our approach is automatic, using machine learning techniques. Thus we overcome the cumbersome task of manually defining heuristics (as some of the related work do), which are limited by the human nature.

At [2] the authors present their work on what they call data frame spreadsheets (i.e., containing attributes or metadata regions on the top/left and a block of numeric values). Using linear-chain, conditional random field (CRF), they perform a sequential classification of rows in the worksheet, in order to infer its layout. Their next immediate focus is extracting the hierarchies found on the top (Header) and left (Attribute) regions. This aspect of their work is further extended at [11]. Additionally, in their first paper, they have experimented with the extraction of the data in the form of relational tuples. They do this based on the information they inferred about the structure of a data frame.

At [12], the authors present their work on schema extraction for Web tabular data, including spreadsheets. They extensively evaluated various methods for table layout inference, all operating at the row level. The CRF classifier combined with their novel approach for encoding cell features into row features (called "logarithmic binning") achieves the highest scores.

The paper [13] presents work on header and unit inference for spreadsheets. Unlike us, the authors take a more software engineering perspective. They utilize the inferred table structure to identify unit errors. The authors have defined a set of heuristics-based spatial-analysis algorithms, and a framework that allows them to combine the results from these algorithms. Additionally, they have evaluated their approach in two datasets, containing 10 and 12 spreadsheets, respectively.

At [14], the authors present DeExcelerator, a framework which takes as input partially structured documents, including spreadsheets, and automatically transforms them into first normal form relations. For spreadsheets, their approach works based on a set of simple rules and heuristics that resulted from a manual study on real-world examples. Their framework operates on different granularity levels (i.e., row, column, and cell), considering the content, formating, and location of the cell/s. They evaluated their system on a sample of 50 spreadsheets extracted from data.gov, using human judges (10 database students).

6.2 HTML Tables

One of the typical ways to present information (facts) is by organizing data in a tabular format. As a result, the problems of table recognition and layout discovery have been encountered by various research communities. Some of the most recent studies are related to HTML (Web) tables. In [15], decision trees and support vector machines (SVM) are considered to differentiate between genuine and non-genuine Web tables. The authors defined structural, content type, and a word group features. The [16] reports the study of large sample of Web tables, which yielded a taxonomy of table layouts. It also discusses heuristics, which are based on features similar to the paper above, to classify Web tables into the proposed taxonomy. In [17], the authors describe the creation of the Dresden Web Table Corpus, by proposing a classification approach that works on the level of different table layout classes.

7 Conclusions

In conclusion, we have presented an updated version of our work, first introduced at [1]. Via cell classification, we aim at identifying the layout and structure of the data in spreadsheets. We defined five labels (classes), based on literature review and our thorough empirical analysis of spreadsheets coming from various domains. Using our specialized tool we initially annotated a considerable sample of worksheets, and subsequently extracted a big variety of predefined features for each annotated cell. The latter, composes our gold standard, which we used for feature selection and for evaluating classifiers. Our experiments show that with the selected features and a Random Forest classifier we can achieve high overall accuracy.

Moreover, we have devised a strategy to fix some of the incorrect classification. Our aim is to get rid of random noise (misclassifications) that might occur in worksheets where we mainly make correct predictions. We attempt to go beyond the rather simplistic approach we discussed at [1], by proposing a three-step process. Initially we cluster our cells into rectangular regions. Similarly to the cell classification, we characterize these regions with a variety of sophisticated features. Subsequently, we use a classifier to identify regions that only contain misclassified cells. Later, we attempt to predict their true label, using another specialized classifier. Our evaluation shows that we perform well at the misclassification identification task, but not as good for the relabeling task. In the future, we plan to define more features for the latter task, and experiment with other classification algorithms as well.

Acknowledgments. This research has been funded by the European Commission through the Erasmus Mundus Joint Doctorate “Information Technologies for Business Intelligence - Doctoral College” (IT4BI-DC).

References

1. Koci, E., Thiele, M., Romero, O., Lehner, W.: A machine learning approach for layout inference in spreadsheets. In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), KDIR, Porto, Portugal, 9–11 November 2016, vol. 1, pp. 77–88 (2016)
2. Chen, Z., Cafarella, M.: Automatic web spreadsheet data extraction. In: SSW 2013, p. 1. ACM (2013)
3. Barik, T., Lubick, K., Smith, J., Slankas, J., Murphy-Hill, E.: FUSE: a reproducible, extendable, internet-scale corpus of spreadsheets. In: MSR 2015 (2015)
4. Hermans, F., Murphy-Hill, E.: Enron’s spreadsheets and related emails: a dataset and analysis. In: Proceedings of ICSE 2015. IEEE (2015)

5. Fisher, M., Rothermel, G.: The EUSES spreadsheet corpus: a shared resource for supporting experimentation with spreadsheet dependability mechanisms. In: SIGSOFT 2005, vol. 30, pp. 1–5. ACM (2005)
6. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth, Belmont (1984)
7. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., Boston (1993)
8. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
9. Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer Series in Statistics. Springer-Verlag New York, Inc., New York (1982)
10. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods - Support Vector Learning. MIT Press (1998)
11. Chen, Z., Cafarella, M.: Integrating spreadsheet data via accurate and low-effort extraction. In: SIGKDD 2014, pp. 1126–1135. ACM (2014)
12. Adelfio, M.D., Samet, H.: Schema extraction for tabular data on the web. In: VLDB 2013, vol. 6, pp. 421–432 (2013)
13. Abraham, R., Erwig, M.: Header and unit inference for spreadsheets through spatial analyses. In: VL/HCC 2004, pp. 165–172. IEEE (2004)
14. Eberius, J., Werner, C., Thiele, M., Braunschweig, K., Dannecker, L., Lehner, W.: DeExcelerator: a framework for extracting relational data from partially structured documents. In: CIKM 2013, pp. 2477–2480. ACM (2013)
15. Wang, Y., Hu, J.: A machine learning based approach for table detection on the web. In: WWW 2002, pp. 242–250. ACM (2002)
16. Crestan, E., Pantel, P.: Web-scale table census and classification. In: WSDM 2011, pp. 545–554. ACM (2011)
17. Eberius, J., Braunschweig, K., Hentsch, M., Thiele, M., Ahmadov, A., Lehner, W.: Building the dresden web table corpus: a classification approach. In: BDC 2015. IEEE/ACM (2015)

Computing Data Lineage and Business Semantics for Data Warehouse

Kalle Tomingas(✉), Priit Järv, and Tanel Tammet

Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia
kalle.tomingas@gmail.com, priit.jarv@gmail.com,
tanel.tammet@gmail.com

Abstract. We present and validate a method and underlying set of technologies, data structures and algorithms to calculate, categorize and visualize component dependencies, data lineage and business semantics from the database structures and queries, independently of actual data in the data warehouse. Chosen approach based on semantic techniques, probabilistic weight calculation and estimation of the impact of data in queries and implemented rule system supports the calculation of the dependency graph from these estimates. We demonstrate a method for business semantics integration and ontology learning from data structures and schemas with a combination of query semantics captured by dependency graph. Annotation of technical assets using a business ontology provides meaning and governance view for human and machine agents to address various planning, automation and decision support problems. Data processing performance and business ontology integration is evaluated and analyzed over several real-life datasets.

Keywords: Data warehouse · Data lineage · Dependency analysis
Data flow visualization · Business semantics · Business ontology

1 Introduction

System developers and managers are facing similar data lineage and impact analysis problems in complex data integration, business intelligence and data warehouse environments where the chains of data transformations are long and the complexity of structural changes is high. The management of data integration processes becomes unpredictable and the costs of changes can be very high due to the lack of information about data flows and the internal relations of system components. Important contextual relations are encoded into data transformation queries and programs (SQL queries, data loading scripts, etc.). Data lineage dependencies are spread between different systems and frequently exist only in program code or SQL queries. Integrating business terminology with technical assets is also challenging for large organizations, especially so in data warehouse and business intelligence contexts. Multiple and heterogeneous data sources, missing schema information, complex data flows, different domains and naming standards are just some of the technical complexities to integrate different technical assets into one semantic model.

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 101–124, 2019.
https://doi.org/10.1007/978-3-319-99701-8_5

All of this leads to unmanageable complexity, lack of knowledge and a large amount of technical work with uncomfortable consequences like unpredictable results, wrong estimations, rigid administrative and development processes, high cost, lack of flexibility and lack of trust.

We point out some of the most important and common questions for large DW which usually become a topic of research for system analysts and administrators:

- Where does the data come or go to in/from a specific column, table, view or report?
- How can we map the technical names to a business glossary?
- When was the data loaded, updated or calculated in a specific column, table, view or report?
- Which components (reports, queries, loadings and structures) are impacted when other components are changed?
- Which data, structure or report is used by whom and when?
- What is the cost of making changes?
- What will break when we change something?

The ability to find ad-hoc answers to many day to day questions determines not only the management capabilities and the cost of the system, but also the price and flexibility of making changes.

The goal of our research is to develop reliable and efficient methods for automatic discovery of component dependencies, data lineage and mappings to business glossary from the database schemas, queries and data transformation components by automated analysis of actual program code. This requires probabilistic estimation of the measure of dependencies and the aggregation and visualization of the estimations.

2 Related Work

Impact analysis, traceability and data lineage issues are not new. An overview of the data lineage and data provenance tracing studies were collected by Cheney et al. [1], historical and future perspectives were discussed by Tan [2] and the last decade of research activities were presented by Pribe et al. [3]. Lineage and provenance has been studied in scientific data processing areas [7–9] and in the context of database management systems [2, 7, 8]. Multiple notions of lineage and provenance in database systems have been used to describe relationships between data in the source and in the target: *where* output records came from [9], *why* an output records were produced by inputs [9, 10] and a *how* output record was produced [11]. The query behavior lineage tracking has been used in classical database problems like view update [12] or the expressiveness of update languages [13], and the study of annotation propagation [13, 14] or updates across peer-to-peer systems [15]. The data-driven and data dependent processes and provenance theoretical and practical models described by Deutch et al. [16].

The distinction is made between coarse-grained, or schema-level, provenance tracking [17] and fine-grained-, or data instance-, level tracking [18]. The methods of extracting the lineage are divided into physical (annotation of data by Missier et al.) and logical, where the lineage is derived from the graph of data transformations [19].

We can also find various research approaches and published papers from the early 1990's and later with methodologies for software traceability [20]. The problem of data lineage tracing in data warehousing environments has been formally founded by Cui and Widom [10, 21]. Data lineage or provenance details levels (e.g., coarse-grained vs fine-grained), question types (e.g., why-provenance, how-provenance and where-provenance) and two different calculation approaches (e.g., eager approach vs. lazy approach) have been discussed in multiple papers [2, 22], and formal definitions of the why-provenance have been given by Buneman et al. [9]. Other theoretical works for data lineage tracing can be found in [23] and [24]. Fan and Poulovassilis developed algorithms for deriving affected data items along the transformation pathway [25]. These approaches formalized a way to trace tuples (resp. attribute values) through rather complex transformations, given that the transformations are known on a schema level. This assumption does not often hold in practice. Transformations may be documented in source-to-target matrices (specification lineage) and implemented in ETL tools (implementation lineage). Woodruff and Stonebraker created a solid base for the data-level and operator processing based the fine-grained lineage, in contrast to the metadata-based lineage calculation in their research paper [26].

Priebe et al. concentrated on proper handling of specification lineage, a significant problem in large-scale DW projects, especially when different sources have to be consistently mapped to the same target [3]. They proposed a business information model (or conceptual business model) as the solution and a central mapping point to overcome those issues. The requirement and design level lineage and traceability solutions for next generation DW and BI architecture described by Dayal et al. [27].

Other ETL-related practical works that are based on conceptual models can be found in [28] and [29]. Ontologies and graphs-based practical works related to data quality and data lineage tracking can be found in [30, 31] and [32]. De Santana proposed the integrated metadata and the CWM metamodel-based data lineage documentation approach [33]. Tomingas et al. employ the Abstract Mapping representation of data transformations and rule-based impact analysis [34]. The conceptual modeling approach of ETL workflows described by Bala et al. [35] in the Big Data landscape and Basal [36] presented a semantic approach to combine the traditional ETL approach with the Big Data challenges. Another related work from the field of data lineage and scientific data provenance by Wang et al. [37] brings together challenges and opportunities of Big Data, including volume, variety, velocity and veracity, with the problems of scientific workflow tracking and reproducibility. The cloud-based or distributed systems have their own limitations for data lineage tracing and the data-centric event logging introduced and discussed by Suen et al. [38].

In addition to data lineage and provenance in databases, closely related workflow provenance tracking is an active research topic in the scientific community. The overview of scientific workflow provenance was captured in surveys by Bose and Frew [6] and Glavic and Dittrich [39], and tutorials with research issues, challenges and opportunities were described by Davidson and Freire in [40].

In the context of our work, efficiently querying the lineage information after the provenance graph has been captured is of specific interest. Heinis and Alonso presented an encoding method that allows space-efficient storage of transitive closure graphs and enables fast lineage queries over that data [17]. Anand et al. proposed a high-level language QLP, together with the evaluation techniques that allow storing provenance graphs in a relational database [41]. These techniques are supported by a pointer-based encoding of the dependency closure that supports reducing storage requirements by eliminating redundancy.

The formal background of ontology engineering, learning and matching are described by Guarino [42, 43]. Ontology learning problems, overview and discussion are provided in a handbook by Maedche and Staab [44], in the context of semantic web by the same authors in [45] and in the context of databases by Li et al. [46] and Astrova [47].

3 Weight Estimation

The inference method of the data flow and the impact dependencies that presented in this paper is part of a larger framework of a full impact analysis solution. The core functions of the system architecture are built upon the following components presented described in detail in our previous works [34, 48]:

1. Scanners collect metadata from different systems that are part of DW data flows (DI/ETL processes, data structures, queries, reports etc.).
2. The SQL parser is based on customized grammars, GoldParser[1] parsing engine and the Java-based XDTL engine.
3. The rule-based parse tree mapper extracts and collects meaningful expressions from the parsed text, using declared combinations of grammar rules and parsed text tokens.
4. The query resolver applies additional rules to expand and resolve all the variables, aliases, sub-query expressions and other SQL syntax structures which encode crucial information for data flow construction.
5. The expression weight calculator applies rules to calculate the meaning of data transformation, join and filter expressions for impact analysis and data flow construction.
6. The probabilistic rule-based reasoning engine propagates and aggregates weighted dependencies.
7. The open-schema relational database using PostgreSQL for storing and sharing scanned, calculated and derived metadata.
8. The directed and weighted sub-graph calculations, and visualization web based UI for data lineage and impact analysis applications.

In the stages preceding the impact estimation, inference and aggregation the data structure transformations are parsed and extracted from queries and stored as formalized, declarative mappings in the system.

To add additional quantitative measures to each column transformation or column usage in the join and filter conditions we evaluate each expression and calculate the transformation and filter weights for those.

[1] http://www.goldparser.org/.

Definition 1. The column transformation weight Wt is based on the similarity of each source column and column transformation expression: the calculated weight expresses the source column transfer rate or strength. The weight is calculated on scale [0, 1] where 0 means that the data is not transformed from source (e.g. constant assignment in a query) and 1 means that the source is copied to the target directly, i.e. no additional column transformations are detected.

Definition 2. The column filter weight Wf is based on the similarity of each filter column in the filter expression where the calculated weight expresses the column filtering strength. The weight is calculated on the scale [0, 1] where 0 means that the column is not used in the filter and 1 means that the column is directly used in the filter predicate, i.e. no additional expressions are involved.

The general column weight W algorithm in each expression for Wt and Wf components is calculated as a column count ratio over all the expression component counts (e.g. column count, constant count, function count, predicate count).

$$W = \frac{IdCount}{IdCount + FncCount + StrCount + NbrCount + PrdCount}$$

The counts are normalized using the FncList evaluation over a positive function list (e.g. CAST, ROUND, COALESCE, TRIM etc.). If the FncList member is in a positive function list, then the normalization function reduces the according component count by 1 to pay a smaller price in case the function used does not have a significant impact to column data.

Definition 3. A primitive data transformation operation is a data transformation between a source column X and a target column Y in a transformation set M (mapping or query) having the expression similarity weight Wt.

Definition 4. The column X is a filter condition in a transformation set M with the filter weight Wf if the column is part of a JOIN clause or WHERE clause in the queries corresponding to M.

4 Rule System and Dependency Calculation

The primitive transformations captured from the source databases form a graph G_O with nodes N representing database objects and edges E_O representing primitive transformations (see Definition 3). We define relations $X : E_O \rightarrow N$ and $Y : E_O \rightarrow N$ connecting edges to source nodes and target nodes, respectively. We define label relations $M : E_O \rightarrow \{\{m\}|m\ \textit{is a transformation identifier}\}$ and $W : E_O \rightarrow [0, 1]$. Formally, this graph is an edge-labeled directed multigraph.

In the remainder of the article, we will use the following intuitive notation: $e.X$ and $e.Y$ to denote source and target objects of a transformation (formally, $X(e)$ and $Y(e)$). $e.M$ is the set of source transformations ($M(e)$). $e.W$ is the weight assigned to the edge ($W(e)$).

The knowledge inferred from the primitive transformations forms a graph $G_L = (N, E_L)$ where E_L is the set of edges e that represent data flow (lineage). We define relations X, Y, M and W the same way as with the graph G_O and use the $e.R$ notation where R is one of the relations $\{X, Y, M, W\}$.

Additionally, we designate the graph $G_I = (N, E_I \cup E_L)$ to represent the impact relations between database components. It is a superset of G_L where E_L is the set of additional edges inferred from column usage in filter expressions.

4.1 The Propagation Rule System

First, we define the rule to map the primitive data transformations to our knowledge base. This rule includes aggregation of multiple edges between pairs of nodes. Let $E_{x,y} = \{e \in E_O | e.X = x, e.Y = y\}$ be the set of edges connecting nodes x, y in the graph G_O.

$$\forall x, y \in N\left\{E_{x,y} \neq \emptyset \Rightarrow \exists e' \in E_L\right\}. \tag{R1}$$

such that

$$e'.X = x \bigwedge e'.Y = y. \tag{R1.1}$$

$$e'.M = \cup_{e \in E_{x,y}} e.M. \tag{R1.2}$$

$$e'.W = max\{e.W | e \in E_{x,y}\}. \tag{R1.3}$$

An inference using this rule (R1) should be understood as ensuring that our knowledge base satisfies the rule. From an algorithmic perspective, we create edges *e'* into the set E_L based on definitions (R1.1–R1.3) until R1 is satisfied.

Definition 5. The predicate *Parent(x, p)* is true if node *p* is the parent of node *x* in the database schema.

Filter conditions are mapped to edges in the impact graph G_I. Let $F_{M,p} = \{x | Parent(x, p) \bigwedge x \textit{ is a filter in } M\}$ be the set of nodes that are filter conditions for the mapping M with parent p. Let $T_{M,p'} = \{x | Parent(x, p') \wedge x \textit{ is target inM}\}$ be the set of nodes that represent the target columns of mapping M. To assign filter weights to columns, we define the function $W_f : N \rightarrow [0, 1]$.

$$\forall p, p' \in N\left\{F_{M,p} \neq \emptyset \bigwedge T_{M,p'} \neq \emptyset \Rightarrow \exists e' \in E_I\right\}. \tag{R2}$$

such that

$$e'.X = p \bigwedge e'.Y = p'. \tag{R2.1}$$

$$e'.M = M. \tag{R2.2}$$

$$e'.W = \frac{max\{W_f(x)|x \in F_{M,p}\} + max\{W_f(x)|x \in T_{M,p'}\}}{2}. \tag{R2.3}$$

The primitive transformations mostly represent column-level (or equivalent) objects that are adjacent in the graph (meaning, they appear in the same transformation or query and we have captured the data flow from one to another). The same applies to impact information inferred from filter conditions. From this knowledge, the goal is to additionally:

- propagate information through the database structure upwards, to view data flows on a more abstract level (such as, table or schema level)
- calculate the dependency closure to answer lineage queries

Unless otherwise stated, we treat the graphs G_L and G_I similarly from this point. It is implied that the described computations are performed on both of them. The set E refers to the edges of either of those graphs. Let $E_{p,p'} = \{\ e \in E|Parent(e.X,p) \bigwedge Parent(e.Y,p')\}$ be the set of edges where the source nodes share a common parent p and the target nodes share a common parent p'.

$$\forall p,p' \in N\{E_{p,p'} \neq \emptyset \Rightarrow \exists e' \in E\}. \tag{R3}$$

such that

$$e'.X = p \bigwedge e'.Y = p'. \tag{R3.1}$$

$$e'.M = \cup_{e \in E_{p,p'}} e.M. \tag{R3.2}$$

$$e'.W = \frac{\sum_{e \in E_{p,p'}} e.W}{|E_{p,p'}|}. \tag{R3.3}$$

4.2 The Dependency Closure

Online queries from the dataset require finding the data lineage of a database item without long computation times. For displaying both the lineage and impact information, we require that all paths through the directed graph that include a selected component are found. These paths form a connected subgraph. Further manipulation (see Sect. 4.3) and data display is then performed on this subgraph.

There are two principal techniques for retrieving paths through a node [17]:

- connect the edges recursively, forming the paths at query time. This has no additional storage requirements, but is computationally expensive
- store the paths in materialized form. The paths can then be retrieved without recursion, which speeds up the queries, but the materialized transitive closure may be expensive to store.

Several compromise solutions that seek to both efficiently store and query the data have been published [17, 41]. In general, the transitive closure is stored in a space efficient encoding that can be expanded quickly at the query time.

We have incorporated elements from the pointer based technique introduced in [41]. The full transitive dependency closure is stored as the union of the pointers to all of the immediate dependency sets of nodes along the paths leading to a selected node.

We can define the dependency closure recursively as follows. Let D_k^* be the dependency closure of node k. Let D_k be the set of immediate dependencies such that $D_k = \{j|e \in E, e.X = j, e.Y = k\}$.

If $D_k = \emptyset$ then $D_k^* = \emptyset$.

Else if $D_k \neq \emptyset$ then $D_k^* = D_k \cup \left(\cup_{j \in D_k} D_j^* \right)$.

The successors S_j (including non-immediate) of a node j are found as follows: $S_j = \{k|j \in D_k^*\}$.

The materialized storage of the dependency closure allows building the successor set cheaply, so it does not need to be stored in advance. Together with the dependency closure they form the connected maximal subgraph that includes the selected node.

We put the emphasis on the fast computation of the dependency closure with the requirement that the lineage graph is sparse ($|I| \sim |N|$). We have omitted the more time-consuming redundant subset and subsequence detection techniques of Anand et al. [49]. The subset reduction has $O(|D|^3)$ time complexity which is prohibitively expensive if the number of initial unique dependency sets $|D|$ is on the order of 10^5 as is the case in our real world dataset.

The dependency closure is computed by:

1. Creating a partial order L of the nodes in the directed graph G_I. If the graph is cyclic then we need to transform it to a DAG by deleting an edge from each cycle. This approach is viable, if the graph contains relatively few cycles. The information lost by deleting the edges can be restored at a later stage, but this process is more expensive than computing the closure on a DAG.
2. Creating the immediate dependency sets for each node using the duplicate-set reduction algorithm [49].
3. Building the dependency closures for each node using the partial order L, ensuring that the dependency sets are available when they are needed for inclusion in the dependency closures of successor nodes.
4. If needed, restoring deleted cyclic edges and incrementally adding dependencies that are carried by those edges using breadth-first search in the direction of the edges.

4.3 Visualization of the Lineage and Impact Graphs

The visualization of the connected subgraph corresponding to a node j is created by fetching the path nodes $P_j = D_j^* \cup S_j$ and the edges along those paths $E_j = \{e \in E|e.X \in P_j \bigwedge e.Y \in P_j\}$ from the appropriate dependency graph (impact or lineage).

The graphical representation allows filtering a subset of nodes in the application, by node type, although the filtering technique discussed here is generic and permits arbitrary criteria. Any nodes not included in graphical display are replaced by transitive edges bypassing these nodes to maintain the connectivity of the dependencies in the displayed graph.

Let $G_j = \left(P_j, E_j\right)$ be the connected sub graph for the selected node *j*. We find the partial transitive graph $G_j^{'}$ that excludes the filtered nodes P_{filt} as follows (Algorithm 1):

Algorithm 1: Building the filtered subgraph with transitive edges.

```
Input: $G_j$, $P_{filt}$
Output: $G_j' = (P_j', E_j')$
$E_j' = E_j$
$P_j' = \varnothing$
for node n in $P_j$:
  if $n \in P_{filt}$:
    for e in $\{e \in E_j' | e.Y = n\}$:
      for e′ in $\{e' \in E_j' | e'.X = n\}$:
        create new edge e′′ (e′′.X = e.X,
                             e′′.Y = e′.Y,
                             e′′.W = e.W * e′′.W)
        $E_j' = E_j' \cup \{e''\}$
      $E_j' = E_j' \setminus \{e\}$
    for e′ in $\{e' \in E_j' | e'.X = n\}$:
      $E_j' = E_j' \setminus \{e'\}$
  else:
    $P_j' = P_j' \cup \{n\}$.
```

This algorithm has the time complexity of O(|Pj| + |Ej|) and can be performed on demand when the user changes the filter settings. This extends to large dependency graphs with the assumption that $|GJ| \ll |G|$.

4.4 The Data Lineage Semantic Layer Calculation

The data lineage semantic layer is a set of visualizations and associated filters to localize the connected subgraph of the expected data flows for the current selected node. All the connected nodes and edges in the semantic layer share the overlapping filter predicate conditions or data production conditions that are extracted during the edge construction. The main idea of the semantic layer is to narrow down all the possible and expected data flows over all the connected graph nodes by cutting down unlikely or disallowed connections in graph, which is based on the additional query filters and the semantic interpretation of filters and calculated transformation expression weights. The semantic layer of the data lineage graph will hide irrelevant and/or highlight the relevant graph nodes and edges, depending on the user choice and interaction.

The visualization of the semantically connected subgraph corresponding to node j is created by fetching the path nodes $P_j = D_j^* \cup S_j$ and the edges along those paths $E_j = \{e \in E | e.X \in P_j \bigwedge e.Y \in P_j\}$ from the appropriate dependency graph (impact or lineage). Any nodes not included in the semantic layer are removed or visually muted (by changing the color or opacity) and the semantically connected subgraph is returned or visualized by the user interface.

Let $G_j = (P_j, E_j)$ be the connected subgraph for the selected node j where $GD_j = (D_j, ED_j)$ is the predecessor subgraph and $GS_j = \left(S_j, ES_j\right)$ is the successor subgraph according to the selected node j. We calculate the data flow graph G_j' that is the union of the semantically connected predecessors $GD_j' = (D_j, ED_j)$ and successor subgraphs $GS_j' = (S_j, ES_j)$. The semantic layer calculation is based on the selected node filter set F_j and calculated separately for back (predecessor) and forward (successors) directions by the recursive algorithm. We will skip the details of the algorithm.

4.5 Dependency Score Calculation

We use the derived dependency graph to solve different business tasks by calculating the selected component(s) lineage or impact over available layers and chosen details. Business questions like: "What reports are using my data?", "Which components should be changed or tested?" or "What is the time and cost of change?" are converted to directed subgraph navigation and calculation tasks. The following definitions add new quantitative measures to each component or node in the calculation. We use those measures in the user interface to sort and select the right components for specific tasks.

Definition 6. Local Lineage Dependency % (LLD) is calculated as the ratio over the sum of the local source and target lineage weights $\mathrm{W_t}$.

$$LLD = \frac{\sum \text{source}(W_t)}{\sum \text{source}(W_t) + \sum \text{target}(W_t)}$$

Local Lineage Dependency 0% means that there are no data sources detected for the object. Local Lineage Dependency 100% means that there are no data consumers (targets) detected for the object. Local Lineage Dependency about 50% means that there are equal numbers of weighted sources and consumers (targets) detected for the object.

Definition 7. Local Impact Dependency % (LID) is calculated as the ratio over the sum of local source and target impact weights $\mathrm{W(W_{t,}, W_f)}$.

$$LLD = \frac{\sum \text{source}(W)}{\sum \text{source}(W) + \sum \text{target}(W)}$$

5 Business Semantics Integration

An integrated semantic model or business ontology would allow us to manage business concepts and the meaning (both human- and machine readable) in a consistent manner and separate form implementation. Annotation of technical assets (e.g. database objects, columns, queries, data transformations tasks, reports, fields, etc.) using a business ontology provides the required meaning, definitions and governance view for human agents and machine readable semantics for machine agents and intelligent applications.

This chapter introduces a novel practical method for business ontology learning from data structures and schemas with a combination of query semantics captured by data lineage graph. We introduce the general dictionary based re-coding of widely used acronyms and naming standards in database schemas. The approach can be easily adapted for domain or dataset specific terminology by adding domain dictionaries.

5.1 Automating Business Terminology Integration

The automation of business terminology integration is based on two complementary data sources:

- schema information of data structures (e.g. a database, etl system or reporting structures and dependencies) and
- calculating the dependency closure of answer lineage queries.

The first provide the basis for terminology, definitions and structure, the latter gives us additional structural and semantic dependencies not available in schema definition, terminology transformations and potential synonyms.

The method of business semantic integration captures and combines information from multiple different data sources. It employs a predefined coding dictionary for technical-to-conceptual translation and a rule based terminology normalization technique for business semantic model generation. The business semantic model is the base for business ontology generation (the bottom-up method) or terms and structure based fuzzy matching of imported business ontology (the top-down method). In other words, the method of business semantic integration is a mixed approach that supports both top down and bottom-up design patterns for ontology engineering. It can be used in an incremental and iterative manner in combination with manual human input and automated mappings. The method enables integration of different data sources and structures into one single corporate business ontology, what is usually the use case for DW and BI environments with multiple connected sources and targets like source databases, ETL, data warehouse, reporting systems and analytical models. It also works in cases when data sources are not necessarily connected and form sparsely connected different domains with their ontologies.

The automation method of business terminology integration is illustrated in Fig. 1 with two complementary data sources S1 and S2 which will be integrated into one semantic model M and can be materialized as a learned business ontology O1 or matched with an imported business ontology O2.

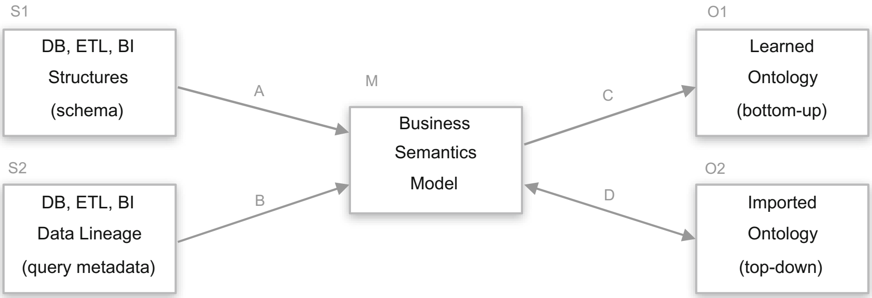


Fig. 1. The automation method of IT assets and business semantics integration.

The data source S1 contains schema information from multiple different databases, data transformation or reporting systems which are not necessarily technically connected to each other.

The data source S2 contains metadata extracted from multiple different data integration and query systems, extracted and stored as a dependency graph described in Sects. 3 and 4 and used as an additional data source to connect assents in S1.

Data items in S1 and S2 transformed with a set of combined techniques A and B to concept candidates and integrated into one business semantics model M. The model M can be materialized or exported as a learned business ontology O1 (the bottom-up approach) with internal structure and asset annotations, or matched with external and imported business ontology O2 (the top-down approach) adding asset annotations and/or new business concept candidates or synonyms to ontology.

The method of A (see Fig. 1) for concept candidate and structure extraction contains techniques of dictionary based recoding and rule based normalization of technical names along with concept type, structure and semantic relation extraction from schema information like table-column structure, primary key and foreign key relations. It is necessary to employ deduplication and aggregation of similar terms. Altogether, the method A contains the following techniques and steps:

- Predefined and/or custom dictionary based concept name normalization, recoding and concept candidate generation;
- Concept type (classes, attributes and instances) detection by data asset types, like tables, views or reports, columns, fields;
- Class attribute and instance concept mapping to sources (tables, views, columns, reports etc.) for annotations;
- Concept candidate definition generation form data asset descriptions and comments;
- Generalization and composition of relation detection by table, view or report structures and PK/FK relations;
- Probabilistic fuzzy duplicate detection and concept matching by name similarity
- Matching score calculation;

The method B (see Fig. 1) uses the additional dependency graph information which contains concept instance and foreign key relation detection using an impact query graph and similar concept/synonym detection using the lineage graph. The method B contains the following techniques and steps in addition to A:

- Foreign key detection using query join and filter conditions from the impact graph for indirect synonymy, association and generalization relation detection;
- Filter predicate and key/value pair detection using query join and filter conditions from the impact graph for instance detection;
- Same data content detection using query transformations from the lineage graph for semantic relations and/or synonym detection and semantic model integration;
- Probabilistic fuzzy concept matching by name similarity, synonyms and a data transformation weight system;
- Matching score calculation;

The dictionary based concept candidate engineering contains a predefined dictionary with the known acronyms and normalized terms widely used in database schemas and technical IT asset definitions, plus unwanted technical acronyms and terms that have no value in a business semantics model. The predefined dictionary can be customized and a domain or dataset specific dictionary can be added to the iterative and incremental process of business semantic model engineering.

The recoding dictionary is loaded to database on every new scan. The extracted, cleaned and normalized terms are added to the dictionary for further manual translation during each iteration in case they are not found in a default standard (e.g. English) dictionary. For example, the predefined dictionary with common acronyms used in a database schema definition would contain mappings like *acct*, *account*; *bal*, *balance; cd, code;* etc.

The name normalization technique for concept candidate generation contains a set of rules like splitting source names by camel-case and a non-alphanumeric token weight system, removal of all non-alphanumeric tokens, matching score calculation, removal of plural form, dictionary re-coding, initial uppercase conversion, short word removal.

The illustrative example of name normalization and re-coding would be following:

- The table name CI_ACCT is converted to a concept candidate **Account**
- The table name CI_ACCT_MSG is converted to a concept candidate **Account Message**
- The column name Acct_Nm is converted to a concept candidate **Account Name**

We will present two connected sample datasets in Figs. 2 and 3 to illustrate the business semantic model generation and ontology learning. The first figure contains database structures from multiple schemas and the second figure is a data lineage graph with data flows between those schema objects.

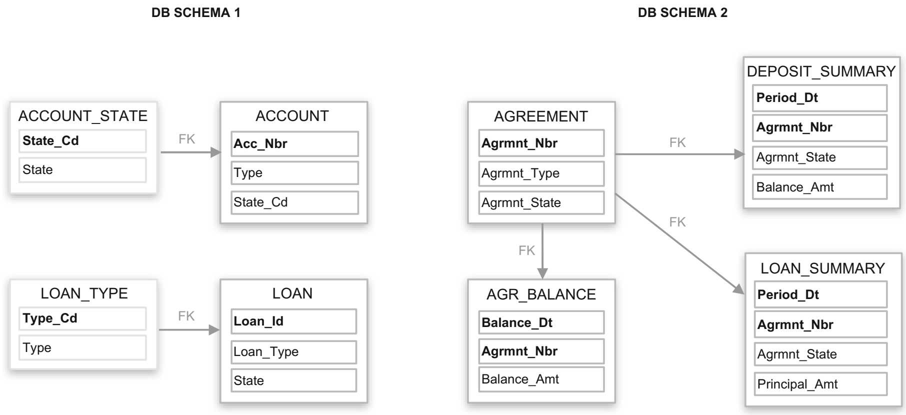


Fig. 2. Sample database schemas and tables.

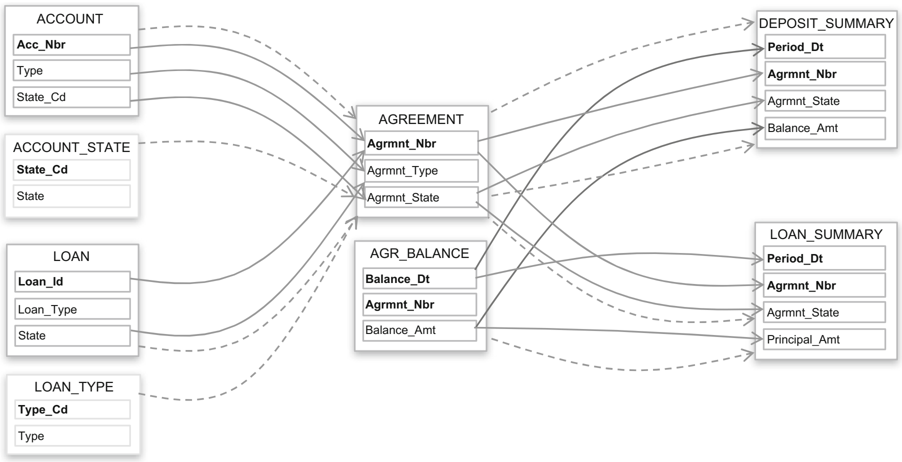


Fig. 3. Sample data flows between tables.

The result of the processing of these two datasets gives us the integrated business semantic model in Fig. 4. which can be exported as a business ontology or matched with an existing imported business ontology. The eight tables and 24 columns (32 items) in an initial data are transformed to 5 classes and 11 attributes (16 concepts) in the business ontology. In this example the 16 concepts cover and annotate 32 DB objects and the number of business definitions is two times smaller than the number of managed assets.

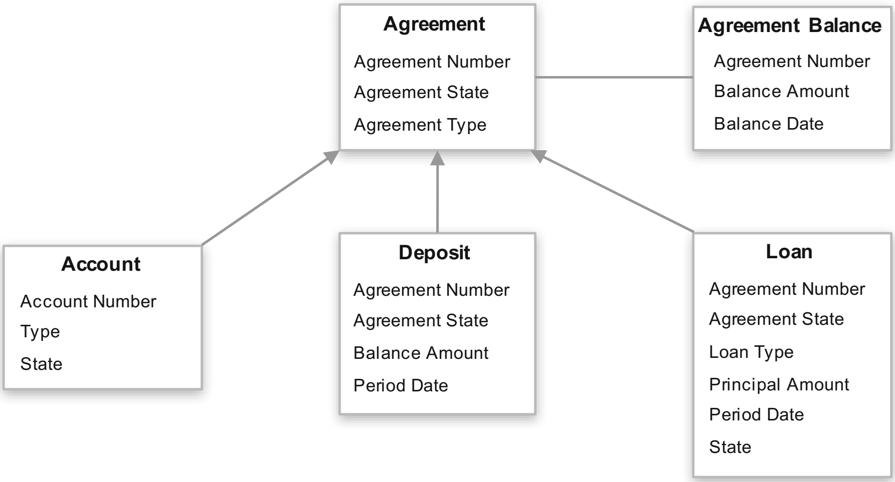


Fig. 4. The ontology learned from database schemas and data lineage metadata.

5.2 Statistics for Semantic Model Learning on Industrial Systems

We will present the relevant statistics of employing the previously described business semantic model learning over six different datasets (Table 1). The datasets DS1 to DS6 represent data warehouse and business intelligence data from different industry sectors and are described in more details in the case study chapter in Sect. 6.1. The structure and integrity of these datasets is diverse. The source structures vary from database schemas to data integration and reporting structures.

Table 1. Business ontology learning evaluation based on six different datasets.

	DS1	DS2	DS3	DS4	DS5	DS6
Nbr of scanned DB, ETL, BI objects	130 171	155 640	39 056	40 545	15 609	2 303
Number of created concepts	**24 195**	**17 168**	**765**	**3 586**	**1 958**	**112**
Number of created classes	4 828	5 490	194	995	718	46
Number of created attributes	19 367	11 678	571	2 591	1 240	66
Number of created annotations	**118 016**	**115 385**	**5 395**	**10 491**	**8 729**	**681**
Concept reuse score (%)	**488**	**672**	**705**	**293**	**446**	**608**
Concept model integrity score (%)	88.8	80.3	22.9	87.5	80.1	22.3
Concept definitions score (%)	0.0	8.0	3.0	7.0	9.0	24.0
Total processing time (min)	**18.1**	**17.1**	**4.1**	**4.3**	**1.8**	**0.6**

The results can be measured in terms of created concepts (classes and attributes) compared to automatically annotated data assets. The concept reuse score measures the converge of semantics and is calculated as a ratio of annotated assets over the number of created concepts. We also calculate the semantic model integrity to measure the connectedness of derived concepts and the definitions score to measure how many concepts are covered by definitions from technical asset descriptions and comments.

6 Case Studies

The previously described algorithms have been used to implement an integrated toolset. Both the scanners and the visualization tools have been enhanced and tested in real-life projects and environments to support several popular data warehouse platforms (e.g. Oracle, Greenplum, Teradata, Vertica, PostgreSQL, MsSQL, Sybase), ETL tools (e.g. Informatica, Pentaho, Oracle Data Integrator, SSIS, SQL scripts and different data loading utilities) and business intelligence tools (e.g. SAP Business Objects, Microstrategy, SSRS). The dynamic visualization and graph navigation tools are implemented in Javascript using the d3.js graphics libraries.

The current implementation has a rule system which is implemented in PostgreSQL database using SQL queries for graph calculation (rules 1–3 in Sect. 4.1) and specialized tables for graph storage. The DB and UI interaction has been tested with both the specialized pre-calculated model (see Sect. 4.2) and the recursive queries without special storage and pre calculations. The algorithms for interactive transitive calculations (see Sects. 4.3) and semantic layer calculation (see Sect. 4.4) are implemented in Javascript and work in a browser for small and local subgraph optimization or visualization. Due to space limitations we will not discuss the details of these case studies. Technical details and additional information can be found on our dLineage[2] online demo site. We will present processing and performance analysis for different datasets in the next section and illustrate the application and algorithms with the graph visualizations technique (Sect. 6.2).

6.1 Performance Evaluation

We have tested our solution in several real-life case studies involving a thorough analysis of large international companies in the financial, utilities, governance, telecom and healthcare sectors. The case studies analyzed thousands of database tables and views, tens of thousands of data loading scripts and BI reports. Those figures are far over the capacity limits of human analysts not assisted by the special tools and technologies.

The following six different datasets with varying sizes have been used for our system performance evaluation. The datasets DS1 to DS6 represent data warehouse and business intelligence data from different industry sectors and is aligned according to the dataset size (Table 2). The structure and integrity of the datasets is diverse and complex, hence we have analyzed the results at a more abstract level (e.g. the number of objects and processing time) to evaluate the system performance under different conditions.

The biggest dataset DS1 contained a big set of Informatica ETL package files, a small set of connected DW database objects and no business intelligence data. The next dataset DS2 contained a data warehouse, SQL scripts for ETL loadings and a SAP Business Object for reporting for business intelligence. The DS3 dataset contained a smaller subset of the DW database (MsSql), SSIS ETL loading packages and SSRS

[2] http://www.dlineage.com/.

reporting for business intelligence. The DS4 dataset had a subset of the data warehouse (Oracle) and data transformations in stored procedures (Oracle). The DS5 dataset is a similar but much smaller to DS4 and is based on the Oracle database and stored procedures. The DS6 dataset had a small subset of a data warehouse in Teradata and data loading scripts in the Teradata TPT format.

Table 2. Evaluation of processed datasets with different size, structure and integrity levels.

	DS1	DS2	DS3	DS4	DS5	DS6
Scanned objects	**1,341,863**	**673,071**	**132,588**	**120,239**	**26,026**	**2,369**
DB objects	43,773	179,365	132,054	120,239	26,026	2,324
ETL objects	1,298,090	361,438	534	0	0	45
BI objects	0	132,268	0	0	0	0
Scan time (min)	**114**	**41**	**17**	**33**	**6**	**0**
Parsed scripts	6,541	8,439	7,996	8,977	1184	495
Parsed queries	48,971	13,946	11,215	14,070	1544	635
Parse success rate (%)	96	98	96	92	88	100
Parse/resolve perform.(queries/sec)	3.6	2.5	26.0	12.1	4.1	6.3
Parse/resolve time (min)	**30**	**57**	**5**	**12**	**5**	**1**
Graph nodes	73,350	192,404	24,878	17,930	360	1,930
Graph links	95,418	357,798	24,823	15,933	330	2,629
Graph processing time (min)	36	62	14	15	6	2
Total processing time (min)	**150**	**103**	**31**	**48**	**12**	**2**

The datasets size, internal structure and processing time are visible in Fig. 5 where longer processing time of DS4 is related to very big Oracle stored procedure texts and loading of those to database.

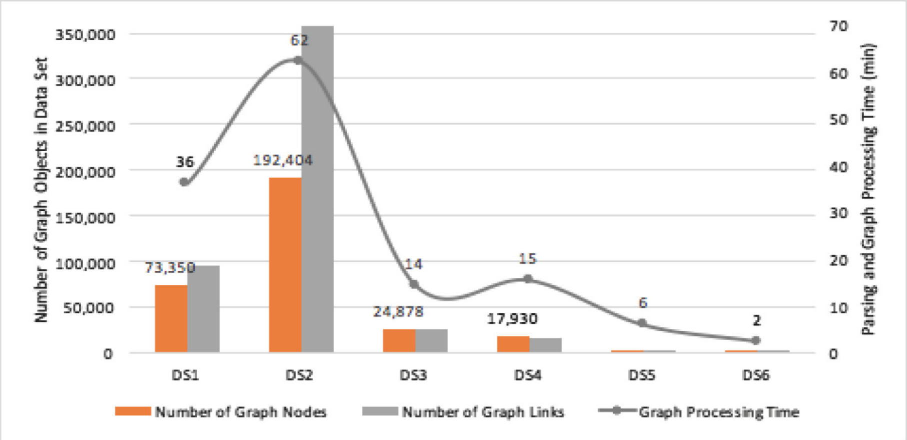


Fig. 5. Datasets size and structure compared to overall processing time.

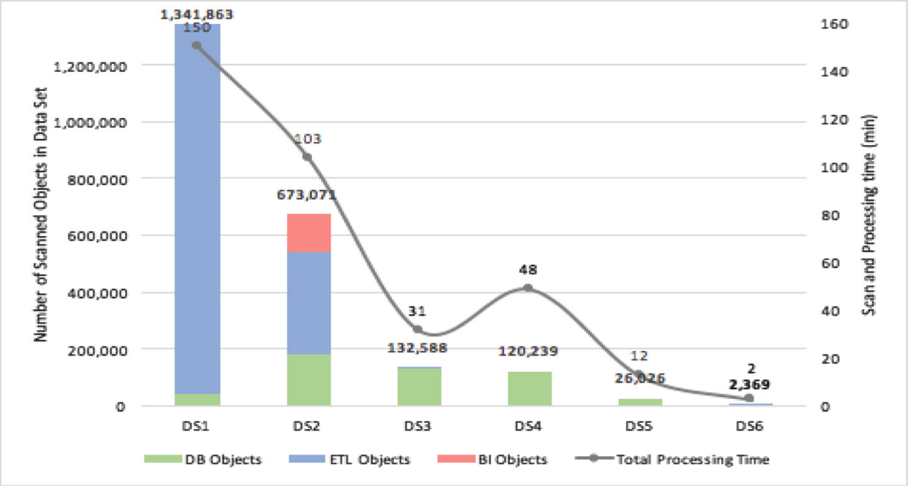


Fig. 6. Calculated graph size and structure compared to the graph data processing time.

The initial dataset and the processed data dependency graphs have different graph structures (see Fig. 6) that do not correspond necessarily to the initial dataset size.

The DS2 has a more integrated graph structure and a higher number of connected objects (Fig. 7) than the DS1. At the same time the DS1 has about two times bigger initial row data size than the DS2.

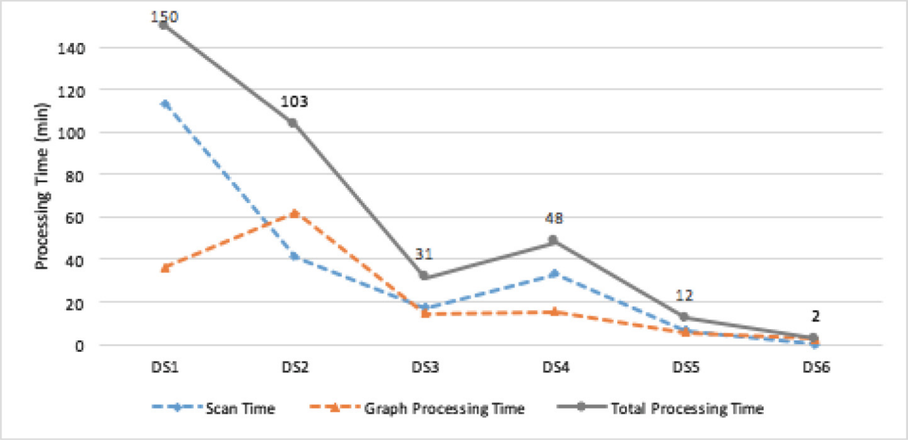


Fig. 7. Dataset processing time with two main sub-components.

We have additionally analyzed the correlation of the processing time and the dataset size (see Fig. 7) and showed that the growth of the execution time follows the same linear trend as the size and complexity growth. The data scan time is related mostly to the initial dataset size. The query parsing, resolving and graph processing time also depends mainly on the initial data size, but also on the calculated graph size (Fig. 7).

6.2 Dataset Visualization

The Enterprise Dependency Graph examples (Figs. 8 and 9) are an illustration of the complex structure of dependencies between the DW storage scheme, access views and user reports. The example is generated using data warehouse and business intelligence lineage layers. The details are at the database and reporting object level, not at column level. At the column and report level the full data lineage graph would be about ten times bigger and too complex to visualize in a single picture. The following graph from the data warehouse structures and user reports presents about 50,000 nodes (tables, views, scripts, queries, reports) and about 200,000 links (data transformations in views and queries) on a single image (see Fig. 8).

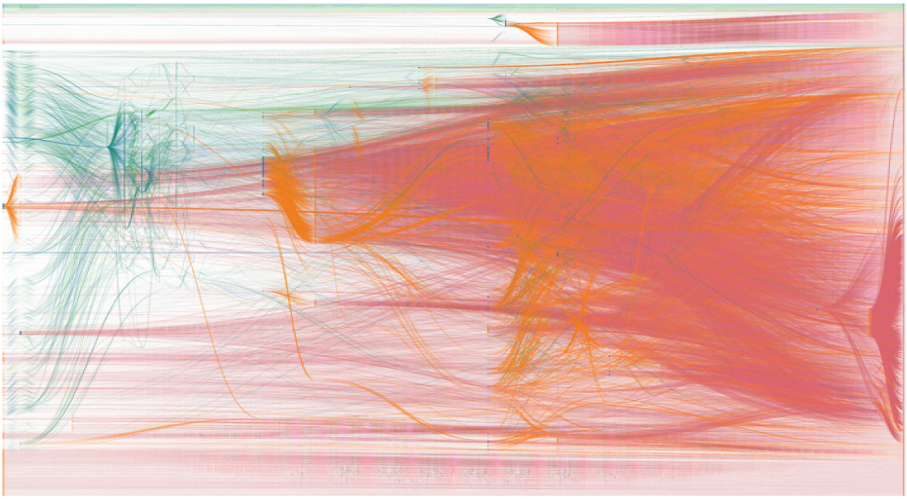


Fig. 8. Data flows (blue, red) and control flows (green, yellow) between tables, views and reports. (Color figure online)

The real-life dependency graph examples illustrate the automated data collection, parsing, resolving, graph calculation and visualization tasks implemented in our system. The system requires only the setup and configuration tasks to be performed manually. The rest will be done by the scanners, parsers and the calculation engine.

The end result consists of data flows and system component dependencies visualized in the navigable and drillable graph or table form. The result can be viewed as a local subgraph with fixed focus and suitable filter set to visualize data lineage path from any sources to single report with click and zoom navigation features. The big picture of

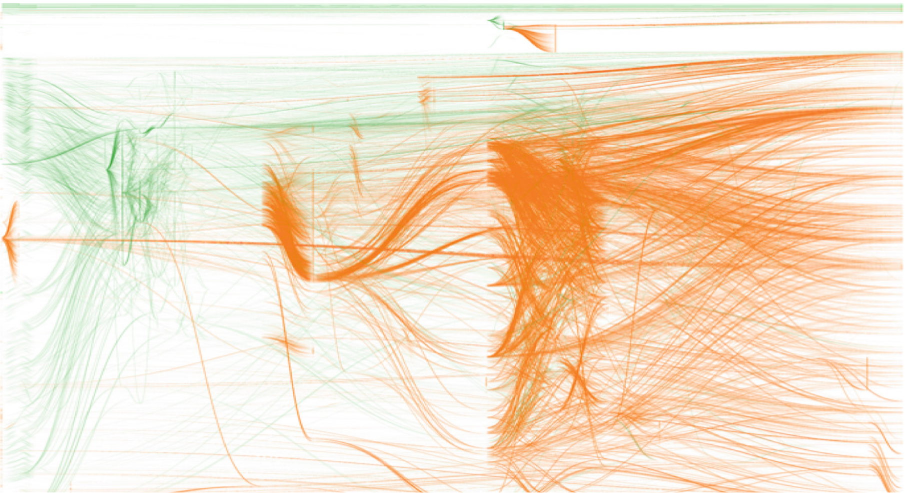


Fig. 9. Control flows in scripts, queries (green) and reporting queries (yellow) are connecting tables, views and reports. (Color figure online)

the dependency network gives the full scale overview graph of the organization's data flows. It allows to see us possible architectural, performance or security problems.

In addition to the visualization of data flows, we developed the aggregated plot view of graph nodes that will help to analyze database tables, data loading programs or reports in terms of connectedness, complexity and cost. The main idea of the visualization is to draw a two-dimensional plot or bubble chart with a number of connected sources and targets on an X and Y axis that allow us to clearly distinguish more and less connected nodes and the balance between the number of sources and targets or data producers and consumers. The size of the bubble in the chart is a recursively calculated number of child objects that express the complexity of the object and its structure. The color of the bubble is calculated as the sum of the all three components – number of sources, targets and children – and it expresses the cost of the object in terms of change, development or maintenance. The more costly objects are in the upper right corner (see Fig. 10), with a bigger diameter and are colored in red. The less costly objects are more in the lower left corner and are colored more in blue. The color layer is the fourth dimension of the chart and it gives a quick and aggregated overview of the selected object set. The bigger and more red an object is, the costlier and more complex it is to change. The smaller and more blue an object is, the less costly and less complex it is to change.

The data axis with its number of sources and targets and bubble size are calculated and drawn in a logarithmic scale to make it more readable. The number of sources, targets and child elements of each object in the same chart can vary with several orders of magnitude, and therefore the logarithmic scale is more suitable for visualization and reading of charts.

The performance evaluation and visualization figures has been partially published in our last work [50].

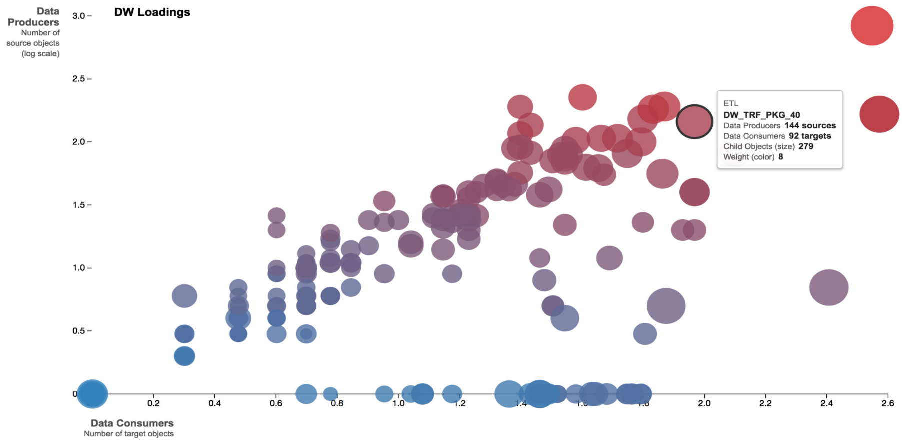


Fig. 10. Data warehouse loading packages with a number of data sources and targets (axis), loading complexity (size) and relative cost (color).

7 Conclusions

We have implemented and presented techniques and algorithms for quantitative impact and data lineage analysis and dependency graph visualizations. The unified data representation and implemented formalized rules allows us to build weighted and directed dependency graphs. Probabilistic weight calculation in query parsing and weight propagation by the rule system brings the data transformation semantics to the graph for further usage. The weight system is also used in the semantic calculation to visualize the applicable data flow subgraphs for each selected node. Integrated business semantic model or business ontology allows us to manage concepts and business meaning separate form implementation. Automated annotation of technical assets using a business ontology provides required meaning, definitions and governance view for human agents and also machine-readable semantics for intelligent agents and applications. The algorithms and techniques have been successfully employed in several large case studies, leading to practical data lineage, component dependency visualizations with integrated business semantics. The presented performance measurements on a number of different big datasets demonstrate the scaling and the computational feasibility of the described approach.

We continue our research and system development in the field of business semantics and governance automation to employ the underlying dependency graph in combination with semantic techniques, ontology engineering and machine learning.

Acknowledgements. The research has been supported by EU through European Regional Development Fund.

References

1. Cheney, J., Chiticariu, L., Tan, W.-C.: Provenance in databases: why, how, and where. Found. Trends Databases **1**(4), 379–474 (2007)
2. Tan, W.: Provenance in databases: past, current, and future. In: SIGMOD 2007, pp. 1–10 (2007)
3. Priebe, T., Reisser, A., Anh Hoang, D.T.: Reinventing the wheel?! Why harmonization and reuse fail in complex data warehouse environments and a proposed solution to the problem. In: Proceedings of the 10th International Conference on Wirtschaftsinformatik, pp. 766–775 (2011)
4. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-Science. SIGMOD Rec. **34**(3), 31–36 (2005)
5. Davidson, S.B., Freire, J.: Provenance and scientific workflows. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data - SIGMOD 2008, p. 1345 (2008)
6. Bose, R., Frew, J.: Lineage retrieval for scientific data processing: a survey. ACM Comput. Surv. **37**(1), 1–28 (2005)
7. Buneman, P., Tan, W.: Provenance in databases. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, pp. 1171–1173 (2007)
8. Zdonik, S.B.: Provenance, lineage, and workflows. In: Computer (Long. Beach. Calif), pp. 1–24 (2010)
9. Buneman, P., Khanna, S., Wang-Chiew, T.: Why and where: a characterization of data provenance. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 316–330. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44503-X_20
10. Cui, Y., Widom, J., Wiener, J.L.: Tracing the lineage of view data in a warehousing environment. ACM Trans. Database Syst. **25**(2), 179–227 (2000)
11. Green, T.J., Karvounarakis, G., Tannen, V.: Provenance semirings. In: Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems - Pod. 2007, no. June, p. 31 (2007)
12. Buneman, P., Khanna, S., Tan, W.-C.: On propagation of deletions and annotations through views. In: Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems - Pod. 2002, vol. 2002, no. June, p. 150 (2002)
13. Buneman, P., Cheney, J., Vansummeren, S.: On the expressiveness of implicit provenance in query and update languages. In: Schwentick, T., Suciu, D. (eds.) ICDT 2007. LNCS, vol. 4353, pp. 209–223. Springer, Heidelberg (2006). https://doi.org/10.1007/11965893_15
14. Bhagwat, D., Chiticariu, L., Tan, W.C., Vijayvargiya, G.: An annotation management system for relational databases. VLDB J. **14**(4), 373–396 (2005)
15. Green, T., Karvounarakis, G.: Update exchange with mappings and provenance. In: Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 675–686 (2007)
16. Deutch, D., Moskovitch, Y., Tannen, V.: A provenance framework for data-dependent process analysis. Proc. VLDB Endow. **7**(6), 457–468 (2014)
17. Heinis, T., Alonso, G.: Efficient lineage tracking for scientific workflows. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data - SIGMOD 2008, Section 2, p. 1007 (2008)
18. Missier, P., Belhajjame, K., Zhao, J., Roos, M., Goble, C.: Data lineage model for taverna workflows with lightweight annotation requirements. In: Freire, J., Koop, D., Moreau, L. (eds.) IPAW 2008. LNCS, vol. 5272, pp. 17–30. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89965-5_4

19. Ikeda, R., Das Sarma, A., Widom, J.: Logical provenance in data-oriented workflows? In: Proceedings - International Conference on Data Engineering, pp. 877–888 (2013)
20. Ramesh, B., Jarke, M.: Toward reference models for requirements traceability. IEEE Trans. Softw. Eng. **27**(1), 58–93 (2001)
21. Cui, Y., Widom, J.: Lineage tracing for general data warehouse transformations. VLDB J. **12** (1), 41–58 (2003)
22. Benjelloun, O., Das Sarma, A., Hayworth, C., Widom, J.: An introduction to ULDBs and the Trio system. IEEE Data Eng. Bull. **29**(1), 5–16 (2006)
23. Fan, H., Poulovassilis, A.: Using AutoMed metadata in data warehousing environments. In: Proceedings of the 6th ACM International of the Work. In: Data Warehouse Ol. - Dol. 2003, p. 86 (2003)
24. Giorgini, P., Rizzi, S., Garzetti, M.: A goal-oriented approach to requirement analysis in data warehouses. Decis. Support Syst. **45**(1), 4–21 (2008)
25. Fan, H., Poulovassilis, A.: Using schema transformation pathways for data lineage tracing. In: Jackson, M., Nelson, D., Stirk, S. (eds.) BNCOD 2005. LNCS, vol. 3567, pp. 133–144. Springer, Heidelberg (2005). https://doi.org/10.1007/11511854_11
26. Woodruff, A., Stonebraker, M.: Supporting fine-grained data lineage in a database visualization environment. In: Proceedings of the 13th International Conference on Data Engineering, no. January, pp. 91–102 (1997)
27. Dayal, U., Castellanos, M., Simitsis, A., Wilkinson, K.: Data integration flows for business intelligence. In: Proceedings of the 12th International Conference on Extending Database Technology Advanced Database Technology - EDBT 2009, p. 1 (2009)
28. Simitsis, A., Vassiliadis, P.: A methodology for the conceptual modeling of ETL processes. In: CAiSE Work, pp. 305–316 (2003)
29. Kabiri, A., Chiadmi, D.: A method for modelling and organizing ETL processes. In: 2nd International Conference on Innovative Computing Technology, INTECH 2012, pp. 138–143 (2012)
30. Skoutas, D., Simitsis, A.: Ontology-based conceptual design of ETL processes for both structured and semi-structured data. Int. J. Semant. Web Inf. Syst. **3**, 1–24 (2007)
31. Galhardas, H., Florescu, D., Shasha, D., Simon, E., Saita, C.-A.: Improving data cleaning quality using a data lineage facility. In: DMDW (2001)
32. Widom, J.: Trio: a system for integrated management of data, accuracy, and lineage. In: Proceedings of the 2005 CIDR Conference, pp. 262–276 (2005)
33. DeSantana, A.S., Moura, A.M.C.: Metadata to support transformations and data & metadata lineage in a warehousing environment. In: Proceedings of 6th International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2004, Zaragoza, Spain, vol. 3181, 1–3 September 2004, pp. 249–258 (2004)
34. Tomingas, K., Kliimask, M., Tammet, T.: Data integration patterns for data warehouse automation. In: Bassiliades, N., et al. (eds.) New Trends in Database and Information Systems II. AISC, vol. 312, pp. 41–55. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-10518-5_4
35. Bala, M., Boussaid, O., Alimazighi, Z.: Extracting-transforming-loading modeling approach for big data analytics. Int. J. Decis. Support Syst. Technol. **8**(4), 50–69 (2016)
36. Bansal, S.K.: Towards a semantic extract-transform-load (ETL) framework for big data integration. In: Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014, pp. 522–529 (2014)
37. Wang, J., Crawl, D., Purawat, S., Nguyen, M., Altintas, I.: Big data provenance: challenges, state of the art and opportunities. In: Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015, pp. 2509–2516 (2015)

38. Suen, C.H., Ko, R.K.L., Tan, Y.S., Jagadpramana, P., Lee, B.S.: S2Logger: end-to-end data tracking mechanism for cloud data provenance. In: Proceedings - 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2013 (2013)
39. Glavic, B., Dittrich, K.: Data provenance: a categorization of existing approaches. In: BTW, pp. 227–241 (2007)
40. Davidson, S., Freire, J.: Provenance and scientific workflows: challenges and opportunities. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1–6 (2008)
41. Anand, M.K., Bowers, S., Ludascher, B.: Techniques for efficiently querying scientific workflow provenance graphs. In: International Conference on Extending Database Technology, pp. 287–298 (2010)
42. Guarino, N.: Formal ontology and information systems. In: Proceedings of the first International Conference on FOIS 1998, vol. 46, no. June, pp. 3–15 (1998)
43. Guarino, N.: Semantic matching: formal ontological distinctions for information organization, extraction, and integration. In: Pazienza, M.T. (ed.) SCIE 1997. LNCS, vol. 1299, pp. 139–170. Springer, Heidelberg (1997). https://doi.org/10.1007/3-540-63438-X_8
44. Maedche, A., Staab, S.: Ontology learning. Handb. Ontol. **13**(3), 245–267 (2004)
45. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intell. Syst. **16**, 72–79 (2001)
46. Li, M.L.M., Du, X.-Y., Wang, S.: Learning ontology from relational database. In: 2005 International Conference on Machine Learning and Cybernetics, vol. 6, no. August, pp. 18–21 (2005)
47. Astrova, I.: Rules for mapping SQL relational databases to OWL ontologies. In: Metadata and Semantics, pp. 415–424 (2009)
48. Tomingas, K., Tammet, T., Kliimask, M.: Rule-based impact analysis for enterprise business intelligence. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H., Sioutas, S., Makris, C. (eds.) AIAI 2014. IAICT, vol. 437, pp. 301–309. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44722-2_32
49. Anand, M.K., Bowers, S., McPhillips, T., Ludäscher, B.: Efficient provenance storage over nested data collections. In: Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology EDBT 2009, p. 958 (2009)
50. Tomingas, K., Järv, P., Tammet, T.: Discovering data lineage from data warehouse procedures 1. In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, pp. 101–110 (2016)

Introducing a Vector Space Model to Perform a Proactive Credit Scoring

Roberto Saia(✉) and Salvatore Carta(✉)

Dipartimento di Matematica e Informatica, Università di Cagliari, Cagliari, Italy
{roberto.saia,salvatore}@unica.it

Abstract. Many authoritative studies report how in these last years the consumer credit was up year on year, making it necessary to develop instruments able to assist the financial operators in some crucial tasks. The most important of them is to classify the loan applications as reliable or unreliable, on the basis of the customer information at their disposal. Such instruments of credit scoring allow the operators to reduce the financial losses, and for this reason they play a very important role. However, the design of effective credit scoring models is not an easy task, since it must face some problems, first among them the data imbalance in the model training. This problem arises because the number of default cases is usually much smaller than that of the non-default ones and this kind of distribution worsens the effectiveness of the state-of-the-art approaches used to define these models. This paper proposes a novel Linear Dependence Based (LDB) approach able to build a credit scoring model by using only the past non-default cases, overcoming both the imbalanced class distribution and the cold-start issues. It relies on the concept of linear dependence between the vector representations of the past and new loan applications, evaluating it in the context of a matrix. The experiments, performed by using two real-world datasets with a strong unbalanced distribution of data, show that the proposed approach achieves performance closer or better than that of one of the best state-of-the-art approaches of credit scoring such as random forests, even using only past non-default cases.

Keywords: Business intelligence · Decision support system
Credit scoring · Data mining · Algorithms · Metrics

1 Introduction

A *credit scoring* process is aimed to evaluate, in terms of reliability, a new loan application (from now on simply named as *instance*), and the acceptation or non acceptation of it depends on its result. For this reason, it is clear that there is a direct correlation between the effectiveness of these models and the gains and losses (i.e., loans that have been fully or partially not repaid) of the financial operators [1].

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 125–148, 2019.
https://doi.org/10.1007/978-3-319-99701-8_6

An ideal approach of *credit scoring* should be able to correctly classify the new instances into two classes, *reliable* or *unreliable*, on the basis of the analysis of the past instances. More formally, *credit scoring* is a statistical approach used to evaluate the probability that a loan application leads to a default [2], allowing the operators to know when a loan can be granted to an applicant [3].

The *credit scoring* is also a powerful instrument for the *risk assessment*, since its process involves all the elements that contribute to determine the probability of loss from a customer's default. It offers the opportunity to reduce the cost related to the credit analysis, allowing the financial operators to monitor the credit activities [4] in real-time, decreasing the response time in the credit decisions.

In the context of credit scoring, as well as it happens in similar contexts (e.g., those related to the *fraud detection* [5]), the unbalanced distribution of data that characterizes the source of information used for the model training represents one of the major problems to face [6].

Such problem happens because the negative cases (i.e., loans that have been fully or partially not repaid) are usually fever than the positive ones (i.e., loans that have been fully repaid), a data configuration that worsens the effectiveness of the machine learning approaches [7].

The core idea of this paper is to *represent the instances in a vector space*, and to *define a metric able to evaluate, in this space, the correlation between a new instance and the other previous non-default ones, in order to evaluate its level of reliability.*

The introduced metric evaluates the reliability of a new instance in terms of linear dependence between its vector representation and those of the past non-default instances. Considering that a set of vectors is linearly independent if no vector in it can be defined as a *linear combination*[1] of the other ones, we believe that *the more the vector representation of a new instance is linearly dependent to the vector representations of the previous non-default ones, the more we can consider it as reliable.*

Such evaluation is performed by calculating the determinant of a matrix $M \times N$, where the first $M-1$ rows are the vector representation of the past non-default instances, and the last M row is the vector representation of the new instance to evaluate (i.e., the size of N is given by the number of features that characterize the instances). Considering that the determinant has no mathematical definition for a non-square matrix, we also introduce a criterion that allows us to manage these cases.

The state-of-the-art approach by which we compare our approach to is *Random Forests*, since in most of the cases reported in the literature [8–10] it outperforms the other ones in the credit scoring tasks.

The main scientific contributions given by this paper are listed below:

(i) formalization of the Average of Sub-matrices Determinants (ASD) criterion able to evaluate the linear dependence of the vector representation of

[1] When one of the vectors is a scalar multiple of the other.

an instance in a vector space, also when this gives rise to a non-square matrix that does not allow us to calculate the determinant by following the canonical criteria;

(ii) calculation of the *Reliability Band* β, which provides information about the linear dependence variations in an ASD process that involves only non-default instances;

(iii) definition of the *Linear Dependence Based* (LDB) approach used to evaluate the new instances, which exploits the β information in order to classify them as *reliable* or *unreliable*;

(iv) experimental demonstration that the LDB approach is able to achieve performance closer or better to that of the state-of-the-art approach we compared our approach to (*Random Forest*), although it operates proactively (i.e., without using default cases), facing both the *cold-start* and the imbalanced class distribution problems.

This paper is based on a previous work presented in [11], which has been completely rewritten and extended with the following contributions:

- presentation of a *parameter tuning* study aimed to detect the best placement of the *Reliability Band* β, which experimentally demonstrates that the choice made in our previous work was the best possible;
- presentation of a *feature selection* study aimed to evaluate the importance of each instance feature in the classification process, made by adopting a *general criterion* based on the feature entropy and a *specific criterion* based on the feature influence in terms of *F-score*;
- evaluation of the proposed approach performance in terms of *sensitivity*, in order to measure its ability to recognize the reliable instances, compared to the state-of-the-art approach taken into account (i.e., *Random Forests*);
- extension of the *Background and Related Work* section by adding some new information and references at the state of the art, presenting to readers a quite exhaustive overview of the context taken into account;
- specification of additional details about the adopted datasets (i.e., features description), in order to better understand some processes performed in this paper (e.g, *parameter tuning* and *feature selection*), as well as the experimental results.

The rest of the paper is organized as follows: Sect. 2 discusses the background and related work; Sect. 3 provides a formal notation and defines the faced problem; Sect. 4 describes the implementation of our *credit scoring* approach; Sect. 5 provides details on the experimental environment, the adopted datasets and metrics, as well as on the used strategy and the experimental results; Some concluding remarks and future work are given in Sect. 6.

2 Background and Related Work

Recent literature proposes several classification techniques able to perform *credit scoring* tasks [12], in addition to a considerable number of other studies aimed to evaluate their performance [8], also by taking into account the impact of the involved parameters [13] and the choice of the metrics for the performace evaluation [14].

Two main advantages related to the use of *credit scoring* approaches [15] are: (*i*) the capability to evaluate the potential risk related to a loan application (i.e., when it is reasonable to grant a loan and when it is not); (*ii*) the capability to infer the customer behavior through a credit scoring model, and then the possibility to propose to them targeted financial services. In this paper we take into account only the first one capability.

2.1 Credit Scoring Models

A *credit scoring* process can exploit a large number of state-of-the-art techniques usually used in the statistic and data mining fields [16,17]. Some representative examples are the linear discriminant models [18], the logistic regression models [19], the neural network models [20,21], the k-nearest neighbor models [22], the genetic programming models [23,24], the entropy-based models [25], and the decision tree models [26,27].

Such techniques can also be combined in order to define new hybrid approaches of *credit scoring*, e.g., as it happens in the techniques that use the neural networks and the clustering methods [28], or in the two-stage hybrid modeling procedure with artificial neural networks and multivariate adaptive regression splines [29,30].

2.2 Public Datasets Availability

It should be observed that almost all the works in literature use a small number of datasets (1.9 on average), as underlined in a recent assessment of the credit scoring scenario performed in [8]. It happens due to the scarcity of datasets publicly available, since the financial operators are reluctant to share their business data, and for other reasons mainly related to privacy issues.

The datasets taken into account in this paper, described in Sect. 5.2, are two of the most used datasets in this research field, and they also well represent a common real-world scenario, which is typically characterized by a strong unbalanced distribution of data.

2.3 Imbalanced Class Distribution

The biggest problem that arises during the definition of a *credit scoring* model is the imbalanced class distribution of data [7,31]. It is an issue related to the fact that the data used to train the model present a small number of default cases and a big number of non-default ones, and such distribution of data reduces the performance of the classification techniques [7,9].

This problem leads toward misclassification costs, an aspect largely faced in literature [32], where some possible solutions have been presented. The most common is the introduction of a preprocessing step that performs an *over-sampling* or *under-sampling* of the classes. The results of such operation, in terms of performance, have been studied in [33,34].

In this work we do not perform any class balancing, because we want to evaluate the proposed approach in the context of a typical real-world dataset.

2.4 Cold Start

The *cold start* issue [35,36] occurs when there is not enough information to train a reliable model about a domain. In the *credit scoring* context such scenario appears when the data used to train the model are not representative of all classes of data [37,38] (i.e., default and non-default cases).

This issue affects a large number of contexts, where it is necessary to define a model based on the previous user interactions, e.g., those related to the recommender systems [39–41], where the previous choices of the users (user profiles) are involved, similarly to the *credit scoring* context, where instead the past loan applications of the users are taken into account.

The proposed approach reduces/overcomes the *cold start* issue by using only a class of data (i.e., the non-default cases during the training dataset) in the model definition process.

2.5 Random Forests

Since its formalization [42], *Random Forests* represents one of the best algorithms among those used for the classification tasks, since its performance usually overcomes those of the other state-of-the-art algorithms.

It represents an ensemble learning method for classification and regression that is based on the construction of a number of randomized decision trees during the training phase, inferring the conclusions by averaging the results.

It is able to manage a wide range of prediction problems, without the need to perform complex configurations, since it only requires the tuning of two parameters: the number of trees and the number of attributes used to grow each tree.

2.6 Matrices, Linearity, and Vector Spaces

The concepts of matrix determinant, linearity, and vector spaces, cover a primary role in the context of this paper, because they are used to formalize the proposed similarity metric based on the linear dependence between vectors, as well as to prove its correctness.

The matrix determinant (*det*) is a mathematical function that assigns a number to every square matrix, so its domain is the set of square matrices, and its range is the set of numbers; more formally, we can write that $det : \Re^n \times \ldots \times \Re^n \rightarrow \Re$.

Regardless of the method used to calculate the determinant of a square matrix $N \times N$ (e.g., by the *Leibniz* formula shown in Eq. 1 [11], where *sgn* is the sign function of permutations σ in the permutation group S_N, which returns $+1$ and -1, respectively for even and odd permutations), its value depends on the relation of linear dependence between the vectors that compose the matrix.

$$det \begin{vmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,N} \\ m_{2,1} & m_{2,2} & \dots & m_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ m_{N,1} & m_{N,2} & \dots & m_{N,N} \end{vmatrix} = \sum_{\sigma \in S_N} sgn(\sigma) \prod_{i=1}^{N} m_{i,\sigma_i} \tag{1}$$

The dependence of the N vectors can be evaluated by calculating the determinant of the $N \times N$ matrix composed by placing, one after the other, the n-tuples that express the vectors in a certain base. The vectors in the matrix are independent when the determinant of the matrix is not zero.

A vector space is a mathematical structure made by a collection of vectors that may be added together and multiplied (or, more correctly, scaled) by numbers called scalars, and it is a set that is closed under finite vector addition and scalar multiplication. A vector sub-space is a vector space that represents a subset of some other vector space of higher dimension.

3 Preliminaries

Formal notation and problem statement related to this paper are stated in the following:

3.1 Notation

Given a set of classified instances $T = \{t_1, t_2, \dots, t_N\}$, and a set of fields $F = \{f_1, f_2, \dots, f_X\}$ that compose each t, we denote as $T_+ \subseteq T$ the subset of non-default instances, and as $T_- \subseteq T$ the subset of default ones.

We also denote as $\hat{T} = \{\hat{t}_1, \hat{t}_2, \dots, \hat{t}_M\}$ a set of unclassified instances, and as $E = \{e_1, e_2, \dots, e_M\}$ these instances after the classification process, thus $|\hat{T}| = |E|$.

It should be observed that an instance can belong only to one class $c \in C$, where $C = \{reliable, unreliable\}$.

3.2 Problem Statement

On the basis of the *linear dependence*, measured by calculating the determinant of the matrices composed by the vector representation of the non-default instances in T_+ and that of the unclassified instances in $\hat{T}$, we classify each instance $\hat{t} \in \hat{T}$ as *reliable* or *unreliable*, by exploiting a *Band of Reliability* β, defined on the basis of the proposed *LDB* approach.

Given a function $eval(\hat{t}, \beta)$ used to evaluate the classification performed by exploiting the β information, which returns a boolean value σ ($0 =$ *wrong classification*, $1 =$ *correct classification*), our objective can be formalized as the maximization of the results sum, as shown in Eq. 2 [11].

$$\max_{0 \leq \sigma \leq |\hat{T}|} \sigma = \sum_{m=1}^{|\hat{T}|} eval(\hat{t}_m, \beta) \tag{2}$$

4 Our Approach

The implementation of the proposed approach, as formerly introduced in [11], is carried out through the following four steps:

1. **Data Normalization:** normalization (lossless) of the F values in a range $[0, 1]$, to make homogeneous the range of involved values, regardless of the considered field or dataset;
2. **ASD Definition:** formalization of the *Average of Sub-matrices Determinants* (ASD) criterion, used to evaluate the linear dependence in the context of a square and non-square matrix of vectors;
3. **Reliability Band Calculation:** definition of the *Reliability Band* β, made on the basis of the ASD criterion, to use in the process of evaluation of the level of reliability of the new instances;
4. **Instances Classification:** formalization of the *Linear Dependence Based* (LDB) algorithm able to classify as *reliable* or *unreliable* a set of unevaluated instances, by exploiting the ASD criterion and the β band.

In the following, we provide a detailed description of each of these steps, since we have introduced the high-level architecture of the proposed LDB. It is presented in Fig. 1 [11], where $T_{+}, \hat{T}$, and E, denote, respectively, the set of non-default instances, the set of instances to evaluate, and the set of classified instances at the end of the process (i.e., those in $\hat{T}$).

4.1 Data Normalization

As first step, we normalize all values F in the datasets T and $\hat{T}$ in a range $[0, 1]$. It allows us to make the range of involved values homogeneous, regardless of the considered field or dataset. Such operation could be also useful in order to avoid potential problems during the determinant calculation, e.g., overflow, in case of very large matrices, by using certain software tools [43].

The process of normalization of a generic value $f_x \in F$ related to an instance $t \in T$ is reported in Eq. 3 [11] (the same goes for the set $\hat{T}$). It should be noted that it is a lossless process.

$$f_x = \frac{1}{\sum\limits_{\forall f_x \in T} f_x} \cdot f_x \tag{3}$$

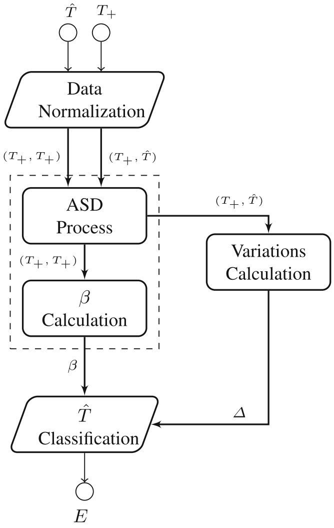


Fig. 1. LDB - high-level architecture.

4.2 ASD Definition

Premising that there is not a mathematical definition of determinant of a non-square matrix, we now introduce the *Average of Sub-matrices Determinants* criterion (ASD), which allows us to extract this information in all cases (square and non-square matrices).

It calculates the average of the determinants of all square sub-matrices obtained by dividing the non-default instance history matrix of size $|T_+| \times |F|$ in α square sub-matrices, whose number depends on the rule shown in Eq. 4 [11].

The additional element added to the set T_+ will be used to evaluate an instance in the context of the vector space of the other instances that composed the matrix.

$$\alpha = \begin{cases} \left\lfloor \dfrac{|F|}{(|T_+|+1)} \right\rfloor, \; if \; |F| \geq (|T_+|+1) \\ \left\lfloor \dfrac{|T_+|+1}{|F|} \right\rfloor, \; otherwise \end{cases} \tag{4}$$

In more detail, we calculate the determinant of each sub-matrix defined by moving (without overlaps) on the matrix $|F| \times (|T_+|+1)$ by using a step $|T_+|+1$ (i.e., along the rows), or by moving on the matrix by using a step $|F|$ (i.e., along the columns). The ASD value is given by the average of all sub-matrices determinant.

By way of example, if the values are $|T_+| = 1$ and $|F| = 6$, we have $\alpha = \left\lfloor \dfrac{6}{1+1} \right\rfloor = 3$ sub-matrices of size 2×2, and the ASD value is calculated as shown in the Eq. 5 [11].

$$ASD\left(\begin{smallmatrix} a & b & c & d & e & f \\ g & h & i & l & m & n \end{smallmatrix}\right) = \frac{det\left(\begin{smallmatrix} a & b \\ g & h \end{smallmatrix}\right) + det\left(\begin{smallmatrix} c & d \\ i & l \end{smallmatrix}\right) + det\left(\begin{smallmatrix} e & f \\ m & n \end{smallmatrix}\right)}{3} \quad (5)$$

From now on, we use the notation $ASD(X, Y)$ to denote the *Average of Submatrices Determinants* calculated by using as the last row of the sub-matrices the vector (or vectors) in the set Y, and for the other rows the vectors in the set X.

Practically, the ASD value gives us information about the linear dependence between vector segments that characterize the same subset of features, as demonstrated in Theorem 1.

Theorem 1. *Given the vector space of the features that characterize the vector representation of transactions in a domain, we can express it as sum of two or more sub-spaces that characterize subsets of features.*

Proof. A vector space can be defined as a combination of sub-spaces by using a decomposition approach, e.g., given a space $\Re^3 = x\text{-}axis + y\text{-}axis + z\text{-}axis$, we can write any $\boldsymbol{w} \in \Re^3$ as a linear combination $c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + c_3\boldsymbol{v}_3$ (where $\boldsymbol{v}$ is a member of the axis, and $c \in \Re$), as shown in Eq. 6 [11].

$$\begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = 1 \cdot \begin{pmatrix} w_1 \\ 0 \\ 0 \end{pmatrix} + 1 \cdot \begin{pmatrix} 0 \\ w_2 \\ 0 \end{pmatrix} + 1 \cdot \begin{pmatrix} 0 \\ 0 \\ w_3 \end{pmatrix} \quad (6)$$

On the basis of the consideration that $\Re^3 = x$-axis$+y$-axis$+z$-axis, we can prove the consistency of the proposed ASD approach, since it gives us the mean value of the determinants calculated on a series of square sub-matrices composed by segments of vectors that belong to the same vector sub-space of the items features space.

It simply means that, by the ASD information, we are able to evaluate subsets of features (in terms of linear dependence between their vector representations), and the calculation of the mean value of these results gives us a single value that reports the relations of similarity in the entire space of the features, as previously demonstrated.

It should also be observed that the previous considerations remain valid in both cases considered by the Eq. 4, because *the determinant of the transpose of any square matrix is the same determinant of the original matrix.*

4.3 Reliability Band Calculation

Denoting as $d(t)$ the ASD value obtained by using as rows of the sub-matrices (except the last row) the non-default instances in T_+, and as last row a vector $t \in T_+$, in Eq. 7 [11] we define the set Δ of ASD variations.

$$\Delta = \{d(t_2) - d(t_1), d(t_3) - d(t_2), \ldots, d(t_N) - d(t_{N-1})\} \quad (7)$$

The *Reliability Band*, denoted as β, is defined by using the average (avg), the minimum (min) and the maximum (max) value of $\varDelta$, as shown in Eq. 8 [11].

$$\beta = [b_L, b_H] = \left[\frac{avg+min}{2}, \frac{avg+max}{2}\right] \tag{8}$$

It gives us information about the linear dependence variations, when the ASD process involves only non-default cases. Such information is used during the evaluation process, by classifying as *unreliable* the cases that generate ASD variations outside the β band (the dotted area shown of Fig. 2 [11]), according to the process explained in the following Sect. 4.4.

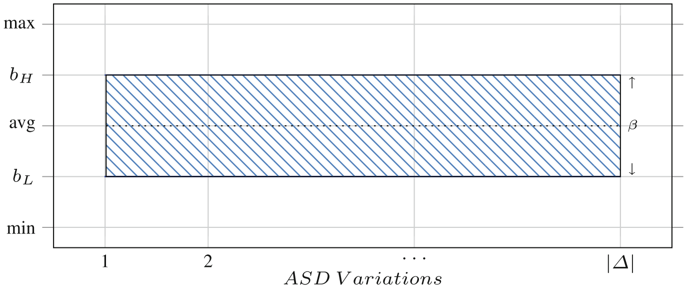


Fig. 2. Reliability band.

4.4 Instances Classification

This section formalizes the algorithm used to perform the *Linear Dependence Based* (LDB) process of classification of an evaluated set of instances, also providing an analysis of its asymptotic time complexity.

Algorithm. The Algorithm 1 takes as input a set T_+ of non-default instances occurred in the past and a set $\hat{T}$ of unevaluated instances, returning as output a set E containing these instances classified as *reliable* or *unreliable* on the basis of the ASD process and the β band (i.e., by using the b_L and b_H values).

In *step 2* we calculate the ASD value by using the non-default instances in T_+ as rows of the sub-matrices (except the last row), and all vectors of the same set as last row (one at a time), as described in Sect. 4.2. The *Reliability Band* β (Sect. 4.3) is calculated in *step 3*. The *steps* from *4* to *11* process the instances $\hat{t} \in \hat{T}$, by using them to fill the last row of each sub-matrix in the ASD process, calculating, in the *step 5*, the variation δ between two instances, following the criterion described in Eq. 7. On the basis of the variation δ and the β band, each instance is classified as *reliable* or *unreliable* in the *steps* from *6* to *10*, and the result is placed in the set E, which is returned at the end of the process (*step 12*).

Algorithm 1. LDB Instances classification.

```
Input: $T_+$=Set of non-default instances, $\hat{T}$=Set of instances to evaluate
Output: $E$=Set of classified instances
 1: procedure InstancesClassification($T_+$,$\hat{T}$)
 2:     $ASD \leftarrow$ getASD($T_+, T_+$)
 3:     $\beta \leftarrow$ getReliabilityBand($ASD$)
 4:     for each $\hat{t}$ in $\hat{T}$ do
 5:         $\delta \leftarrow$ getASD($T_+, \hat{t}_m$) - getASD($T_+, \hat{t}_{m-1}$)
 6:         if $\delta \geq \beta(b_L)$ AND $\delta \leq \beta(b_H)$ then
 7:             $E \leftarrow$ ($\hat{t}$,$reliable$)
 8:         else
 9:             $E \leftarrow$ ($\hat{t}$,$unreliable$)
10:         end if
11:     end for
12:     return $E$
13: end procedure
```

Considering that the calculation of the linear dependence variations (*step 5*) needs at least two instances, when we evaluate the first instance of the set $\hat{T}$ (or when there is only an instance in this set), we add an additional instance composed by using the average of each $f \in F$ of the set T_+, as first instance of the set $\hat{T}$. For algorithm readability reasons, we omitted this step, as well as that of the preliminary normalization of the sets T and $\hat{T}$.

Asymptotic Time Complexity Analysis. Taking into account the possible implementation of the *LDB* approach in a *real-time scoring system* [44], where the *response-time* factor could represent an important element, here we define the theoretical complexity analysis of the Algorithm 1.

We denote as $N = \alpha$ the dimension of the input, since it is related to the number of sub-matrices involved in the ASD process, as shown in Eq. 4. Considering that:

(i) the complexity (*Big O notation*) of the *step 2* is $O(N^2)$, since it performs the ASD process by using N instances for N times, i.e., $ASD(T_+, T_+)$;
(ii) the complexity of the *step 3* is $O(1)$, because it obtains all needed information at the end of the previous *step 2*;
(iii) the complexity of the cycle in the *steps 4–11* is the same of the *step 2*, because it performs the same operation by using the items in the set $\hat{T}$ instead of the ones of the set T_+.

On the basis of the previous considerations, we can define the asymptotic time complexity of the algorithm as $O(N^2)$.

The computational time may be reduced by distributing the process over different machines, by employing large scale distributed computing models (e.g., such as *MapReduce* [45]).

5 Experiments

This section describes the experimental environment, the used datasets and metrics, the adopted strategy (where we perform the *parameter tuning* and the *feature selection* tasks), and the results of the performed experiments.

5.1 Environment

The machine used for the experiments was an Intel i7-4510U, quad core (2 GHz × 4) and a Linux 64-bit Operating System (*Debian Jessie*) with 8 GBytes of RAM. The proposed approach was developed in Java by using the *Java Matrix Package* (*JAMA*)[2].

5.2 Datasets

The two real-world datasets used during the experiments have been chosen for two reasons: first, they represent two benchmarks in this research field; second, both of them are characterized by a strong unbalanced distribution of data. The first one is the *German Credit* (GC) and the *Default of Credit Card Clients* (GC) datasets and the second one is the *Credit Approval* (CA) dataset.

Both the datasets are available at the *UCI Repository of Machine Learning Databases*[3]. They are released with all the attributes modified to protect the confidentiality of the data, and we use a version suitable for algorithms that, as the one proposed, can not operate with categorical variables (i.e., a version with all numeric attributes). It should be noted that, in case of other datasets that contain categorical variables, their conversion in numerical ones is usually a simple task.

Table 1. Datasets overview.

Dataset name	Total cases	Non-default	Default	Attributes	Classes
	$\|T\|$	$\|T_+\|$	$\|T_-\|$	$\|F\|$	$\|C\|$
GC	$1,000$	700	300	21	2
DC	$30,000$	$23,364$	$6,636$	23	2

The datasets' characteristics are summarized in Table 1 [11] and detailed in the following:

German Credit (GC). It contains *1,000* instances: *700* of them are non-default instances (70.00%) and *300* are default instances (30.00%). Each instance is composed by 20 features (whose type is described in Table 2) and a binary class variable (*reliable* or *unreliable*).

[2] http://math.nist.gov/javanumerics/jama/.

[3] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/.

Table 2. Dataset GC features.

Feature	Description	Feature	Description
01	*Status of checking account*	*11*	*Present residence since*
02	*Duration*	*12*	*Property*
03	*Credit history*	*13*	*Age*
04	*Purpose*	*14*	*Other installment plans*
05	*Credit amount*	*15*	*Housing*
06	*Savings account/bonds*	*16*	*Existing credits*
07	*Present employment since*	*17*	*Job*
08	*Installment rate*	*18*	*Maintained people*
09	*Personal status and sex*	*19*	*Telephone*
10	*Other debtors/guarantors*	*20*	*Foreign worker*

Default of Credit Card Clients (DC). It contains *30,000* instances: *23,364* of them are non-default instances (77.88%) and *6,636* are default instances (22.12%). Each instance is composed by 23 features (whose type is described in Table 3) and a binary class variable (*reliable* or *unreliable*).

Table 3. Dataset DC features.

Feature	Description	Feature	Description
01	*Credit amount*	*13*	*Bill statement in Aug-2005*
02	*Gender*	*14*	*Bill statement in Jul-2005*
03	*Education*	*15*	*Bill statement in Jun-2005*
04	*Marital status*	*16*	*Bill statement in May-2005*
05	*Age*	*17*	*Bill statement in Apr-2005*
06	*Past repayments in Sep-2005*	*18*	*Amount paid in Sep-2005*
07	*Past repayments in Aug-2005*	*19*	*Amount paid in Aug-2005*
08	*Past repayments in Jul-2005*	*20*	*Amount paid in Jul-2005*
09	*Past repayments in Jun-2005*	*21*	*Amount paid in Jun-2005*
10	*Past repayments in May-2005*	*22*	*Amount paid in May-2005*
11	*Past repayments in Apr-2005*	*23*	*Amount paid in Apr-2005*
12	*Bill statement in Sep-2005*		

5.3 Metrics

This section presents the metrics used in the context of this paper.

Shannon Entropy. The Shannon entropy, formalized by *Claude E. Shannon* in [46], is one of the most important metrics used in information theory. It reports the uncertainty associated with a random variable, allowing us to evaluate the average minimum number of bits needed to encode a string of symbols, based on their frequency.

More formally, given a set of values $v \in V$, the entropy $H(V)$ is defined as shown in the Eq. 9, where $P(v)$ is the probability that the element v is present in the set V.

In the context of the classification tasks, the entropy-based metrics are frequently used for the *feature selection* [47–49] process, which is aimed to detect a subset of relevant features (variables, predictors) to use during the model definition.

$$H(V) = -\sum_{v \in V} P(v) log_2[P(v)] \tag{9}$$

Accuracy. The *Accuracy* metric reports the number of instances correctly classified (i.e., *true positives* plus *true negatives*), compared to the total number of them. It gives us an overview about the classification performance.

Formally, given a set of instances $\hat{T}$ to be classified, it is calculated as shown in Eq. 10 [11], where $|\hat{T}|$ stands for the total number of instances, and $\hat{T}^{(+)}$ stands for the number of those correctly classified.

$$Accuracy(\hat{T}) = \frac{\hat{T}^{(+)}}{|\hat{T}|} \tag{10}$$

Sensitivity. Differently from the *accuracy* metric previously described, which takes into account all kind of classifications, through the *sensitivity* (also known as *true positive rate*) we only obtain information about the number of instances correctly classified as *reliable*. It gives us an important information, since it evaluates the predictive power of our approach in terms of capability to detect the reliable loan applications, and this is why it represents one of the most used metrics for the evaluation of the credit scoring approaches [8].

More formally, given a set of instances X to be classified, the *Sensitivity* is calculated as shown in Eq. 11, where $|X^{(TP)}|$ stands for the number of instances correctly classified as *reliable* and $|X^{(FN)}|$ for the number of *reliable* instances wrongly classified as *unreliable*.

$$Sensitivity(X) = \frac{|X^{(TP)}|}{|X^{(TP)}|+|X^{(FN)}|} \tag{11}$$

F-Score. The *F-score* [50] is the weighted average of the *precision* and *recall* metrics. It is a largely used metric in the statistical analysis of binary classification and gives us a value in a range $[0, 1]$, where 0 represents the worst value and 1 the best one.

Formally, given two sets X and Y, where X denotes the set of performed classifications of instances, and Y the set that contains the actual classifications of them, this metric is defined as shown in Eq. 12 [11].

$$\begin{aligned} &F\text{-}score(X,Y) = 2\cdot\frac{(precision(X,Y)\cdot recall(X,Y))}{(precision(X,Y)+recall(X,Y))}\\ &\text{with}\\ &precision(X,Y) = \frac{|Y \cap X|}{|X|},\ \ recall(X,Y) = \frac{|Y \cap X|}{|Y|} \end{aligned} \tag{12}$$

AUC. The *Area Under the Receiver Operating Characteristic* curve (*AUC*) is a performance measure [50,51] used to evaluate a predictive model of *credit scoring*. Its result is in a range $[0, 1]$, where 1 indicates the best performance.

Given the subset T_+ of non-default instances in the set T and the subset T_- of default ones, all possible comparisons Θ of the scores of each instance t are reported in the Eq. 13 [11], and the *AUC* metric, by averaging over these comparisons, can be written as in Eq. 14.

$$\Theta(t_+, t_-) = \begin{cases} 1, & if\ t_+ > t_- \\ 0.5, & if\ t_+ = t_- \\ 0, & if\ t_+ < t_- \end{cases} \tag{13}$$

$$AUC = \frac{1}{T_+ \cdot T_-} \sum_{1}^{|T_+|} \sum_{1}^{|T_-|} \Theta(t_+, t_-) \tag{14}$$

5.4 Strategy

This section reports information about the strategy adopted during the execution of the experiments.

Cross-Validation. In order to minimize the impact of data dependency and improve the reliability of the obtained results [52], the experiments have been performed by following the *k-fold cross-validation* criterion, with $k = 10$: each dataset is randomly shuffled and then divided in k subsets; each subset k is used as test set, while the other $k-1$ subsets are used as training set; at the end of the process, we consider the average of results.

Parameter Tuning. Before starting the experiments we carried out a study aimed to identify the best placement of the reliability band β in the range between the minimum (*min*) and maximum (*max*) values of the *ASD variations*, as described in Sect. 4.3 and shown in Fig. 2. In more detail, we translate the reliability band β from the minimum value of the *ASD variations* (i.e., its bottom coincides with the *min* value of Fig. 2) to the maximum value of it (i.e., its top coincides with the *max* value of Fig. 2).

We perform this operation by subdividing the translation in *20* steps, where step *1* denotes the lowest position of the reliability band and step *20* the highest position of it. It means that step *10* denotes the canonical position of β (i.e., when it is centered in the average value of the *ASD variations*, as shown in Fig. 2).

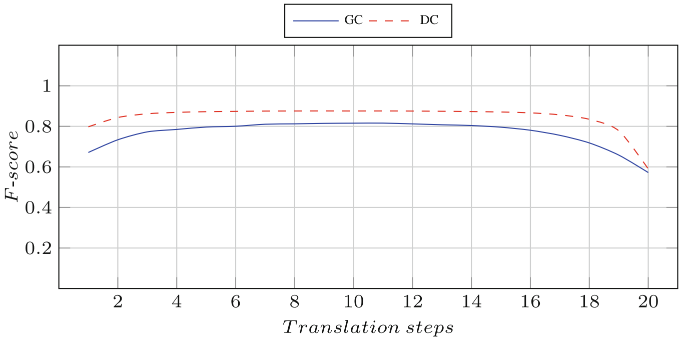


Fig. 3. Reliability band tuning.

The results of Fig. 3 experimentally prove that the best choice is to center the reliability band β on the average value of the *ASD variations*, as we deductively stated in our previous work [11].

Feature Selection. Subsequently to the identification of the best placement of the reliability band β, we performed another series of experiments in order to evaluate the contribution of each instance feature in the classification process.

Many studies [53] have discussed how the performance of credit scoring models is strongly related to the features used to define them. Such process is usually defined as *feature selection* and it can be performed by exploiting different techniques, on the basis of the characteristics of the context taken into account.

It means that the choosing of the best features to use during the model definition is not based on a unique criterion but rather it exploits several criteria with the aim to evaluate, as best as possible, the influence of each feature in the process of definition of the credit scoring model.

It represents an important preprocessing step, since it can reduce the complexity of the final model, decreasing the training times, and increasing the generalization of the model. It also can reduce the problem related with the *overfitting*, a problem that occurs when a statistical model describes random error or noise instead of the underlying relationship, and this frequently happens during the definition of excessively complex models, since many parameters, with respect to the number of training data, are involved.

In this paper, we face the aforementioned problem by performing an empirical study based on two criteria: (*i*) we first measure the Shannon entropy (i.e., a metric described in Sect. 5.3) of each feature in order to evaluate its contribution in the instance characterization; (*ii*) we verify the previous entropy-based evaluation in a real-world context, by removing the features, one by one, measuring the variations in terms of *F-score*.

On the basis of the results presented in Fig. 4 we can observe that although several features present a high level of entropy (i.e., a low level of instance characterization, since the entropy increases as the data becomes equally probable),

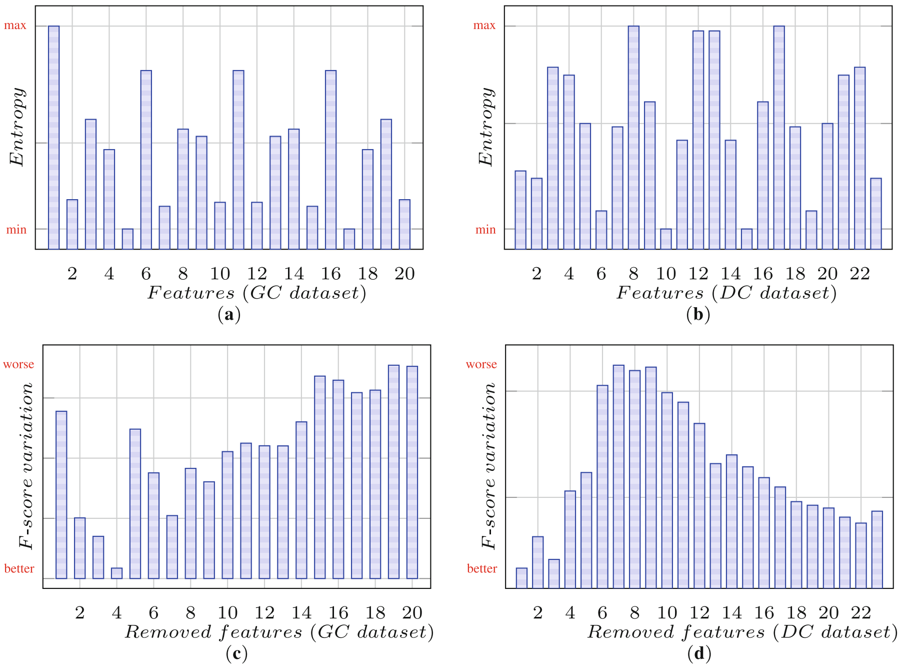


Fig. 4. Instance features entropy and influence.

they have a positive contribute in the classification process, because their removal reduces the performance in terms of *F-score*.

Premising that Fig. 4c and d show the *F-score* variation when a single feature (that in the x-axis) is not taken into account during the model definition process, we observe the following:

(i) although the features *1*, *6*, *11*, and *16* of the *GC* dataset have a high level of entropy (Fig. 4a), if we do not consider them during the model definition process, the *F-score* gets worse (Fig. 4c). As reported in Table 2, these features respectively refer to the *status of existing checking account*, the *savings account/bonds*, the *present residence since*, and the *existing credits*;

(ii) although the features *8*, *12*, *13*, and *17* of the *DC* dataset have a high level of entropy (Fig. 4b), if we do not consider it during the model definition process, the *F-score* gets worse (Fig. 4d). As reported in Table 3, these features respectively refer to the *past repayment in Jul-2005*, the *Bill statement in Sep-2005*, the *Bill statement in Aug-2005*, and the *Bill statement in Apr-2005*;

(iii) the *F-score* evaluation, made by removing the features one by one, shows that the minor contribution in the model definition process is given by the features *2*, *3*, and *4* of the *GC* dataset, and by the features *1*, *2*, and *3* of the *DC* dataset, in contrast to the previous entropy-based evaluation.

On the basis of the aforementioned observations, we can deduce that a mere entropy criterion is not able to detect the best features to use in the model definition process. It depends on the fact that some features with a high level of entropy (equiprobability in the prediction of their values) represent crucial information for the credit scoring, as proved by the results shown in Fig. 4c and d.

Table 4. Feature selection results.

Dataset name	Accuracy loss	Sensitivity loss	F-score loss	AUC loss
GC	0.007	0.000	0.003	0.037
DC	0.005	0.009	0.004	0.115

As shown in Table 4, the removal of the features *2*, *3*, and *4* from the instances in the GC dataset, and the removal of the features *1*, *2*, and *3* from the instances in the DC dataset, leads toward a negligible reduction of the performance of our LDB approach.

However, despite its quite lossless impact on the general performance (in terms of *precision*, *sensitivity*, *f-score*, and AUC metrics), this process significantly reduces the computational complexity, since after the *feature selection* we exclude from the model definition process $21,000$ elements of the GC dataset and $70,092$ elements from the DC dataset, on the basis of the respective training sets (i.e., the non default cases in the set T_+). Although it represents a very considerable advantage in the context of more complex real-world scenarios, considering the size of the adopted datasets, we decided to use all instance features during the experiments.

5.5 Competitors

The implementation of the state-of-the-art approach to which we compare our approach was made in R[4], by using the *randomForest* and *ROCR* packages.

For reasons of reproducibility of the RF experiments, we fix the seed of the random number generator by calling the R function *set.seed()*. About the tuning of the RF parameters, they have been defined experimentally, by researching those that maximize the classification performance.

5.6 Results

In this section we present and discuss the experimental results.

[4] https://www.r-project.org/.

Overview. By observing the experimental results reported in the Figs. 5, 6, 7, and 8, it is possible to make the following considerations:

(i) in terms of *Accuracy*, the performance of our *LDB* approach is very close to those of the *RF* one on both the *GC* and *DC* datasets, as shown in Fig. 5;
(ii) in terms of *Sensitivity*, our *LDB* approach gets better results than the *RF* one, on both the *GC* and *DC* datasets, as shown in Fig. 6;
(iii) in terms of *F-score*, our *LDB* approach outperforms the *RF* on the *DC* dataset, and it is very close to it on the *GC* dataset, as shown in Fig. 7;
(iv) in terms of *AUC*, the performance of our *LDB* approach is very close to those of *RF* on both the *GC* and *DC* datasets, as shown in Fig. 8;
(v) there is no dominant subsets of data that influence the overall performance of the *LBD* approach, since the values of the evaluation metrics remains quite stable along all the *10* subsets (i.e., the *k* folds used in the *k-fold cross-validation process*), as shown in Fig. 9.

It should be added that the independent-samples *two-tailed Student's t-tests* highlighted that there is a statistical difference between the results ($p < 0.05$).

Discussion. By observing the experimental results in more detail, we can notice that our *LDB* approach obtains a level of performance very close or better to those of the *RF* one, although it does not use the past default cases during the model training.

Another aspect is related to the *F-measure* results, which underline how the performance of our *LDB* approach is correlated to the number of past non-default instances used in the model training (*DC* dataset). It means that, differently from the *RF* approach, the *LDB* performance improves when the number of past non-default instances increases.

The performance of our *LDB* approach is very interesting also in terms of the *AUC* metric, as reported in Fig. 8. Such metric measures the predictive power of a classification approach, and the results show that the performance of *LDB* is similar to those of *RF*, again considering that that we do not train our model with the past default instances.

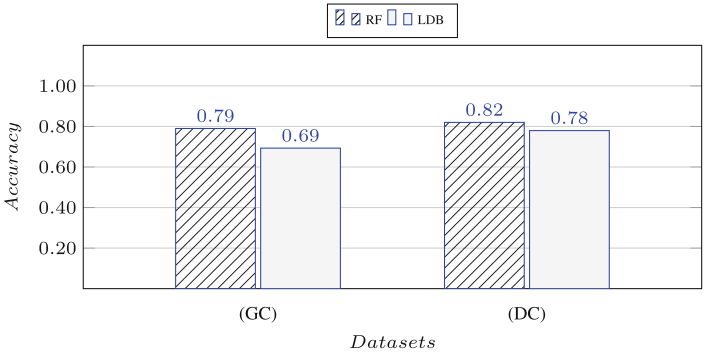


Fig. 5. General performance: accuracy.

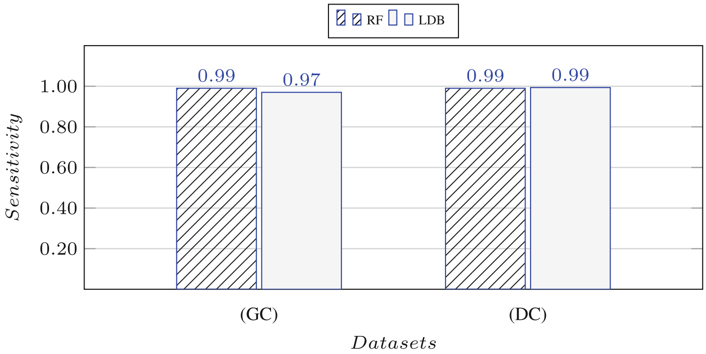


Fig. 6. General performance: sensitivity.

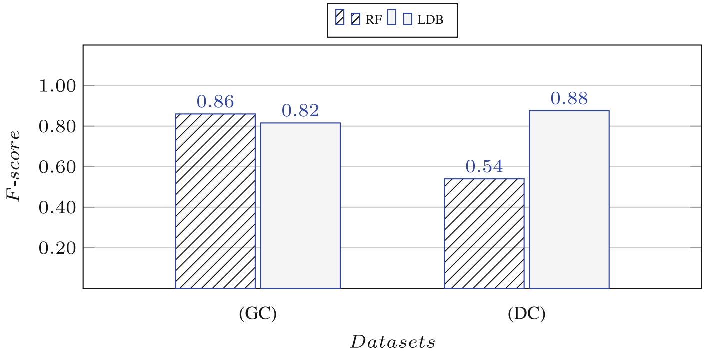


Fig. 7. General performance: F-score.

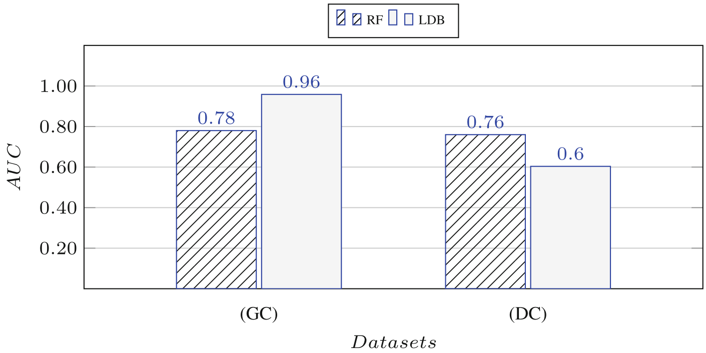


Fig. 8. AUC performance.

According to the previous considerations, we can note that a side effect of our proactive modality is its capability to face the *cold-start* issue previously introduced in Sect. 2.4. It means that it allows us to operate in a real-world context, even when we do not have previous default cases to use in the model training, with all the benefits that derive from it.

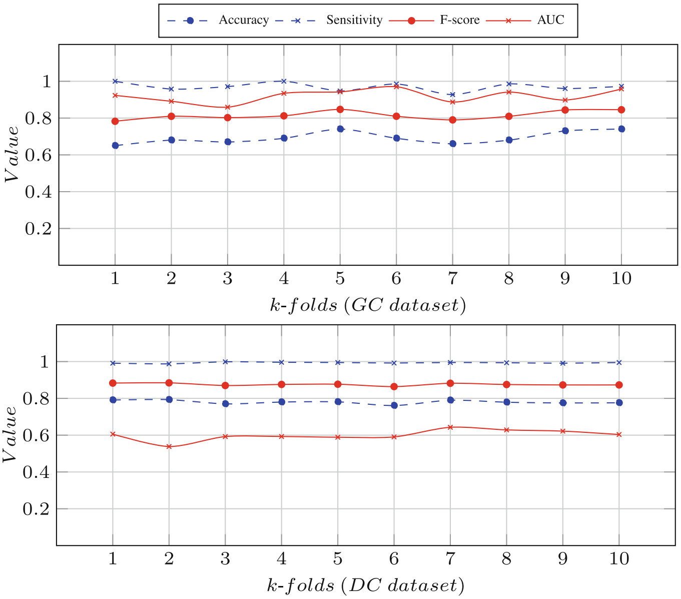


Fig. 9. Performance by k-folds: GC and DC datasets.

6 Conclusions and Future Work

Nowadays, the *credit scoring* techniques became more and more important thanks to their capability to assist the financial operators in many crucial tasks (e.g., bank loans, mortgage lending, insurance policies, etc.).

In the context taken into account in this paper, such techniques are aimed to classify a new loan application as reliable or unreliable (i.e., when a loan can be granted to an applicant), on the basis of the past cases.

It is clear that the effectiveness of these techniques is directly related with the losses due to default, and for this reason there is a continuous research of even more effective *credit scoring* techniques.

In this paper, we proposed a novel *LDB* approach of *credit scoring* based on the concept of *Linear Dependence*, which is used in order to classify the new instances as *reliable* or *unreliable*.

Such approach presents two main advantages: on the one hand it faces the data unbalance issue by involving only a class of data during the model training, giving rise to a proactive modality that allow us to reduce/overcome the *cold-start* problem;

On the other hand, it outperforms the state-of-the-art approach taken into account (i.e., *Random Forests*), when the training process uses a large number of non-default instances, which is an interesting result, considering that it does not exploit both classes of instances (default and non-default cases).

Future work would analyze the performance of our *LDB* approach when we also include the default past cases in the training process, by evaluating the advantages and disadvantages related to the use of such non-proactive strategy.

Another interesting future work would be the evaluation of our *LDB* approach in the context of heterogeneous financial environments, in which many type of data are involved, as it happens in the e-commerce environment.

Acknowledgements. This research is partially funded by Regione Sardegna under project "Next generation Open Mobile Apps Development" (NOMAD), "Pacchetti Integrati di Agevolazione" (PIA) - Industria Artigianato e Servizi - Annualità 2013.

References

1. Henley, W., et al.: Construction of a k-nearest-neighbour credit-scoring system. IMA J. Manag. Math. **8**, 305–321 (1997)
2. Mester, L.J.: What's the point of credit scoring? Bus. Rev. **3**, 3–16 (1997)
3. Morrison, J.: Introduction to survival analysis in business. J. Bus. Forecast. **23**, 18 (2004)
4. Brill, J.: The importance of credit scoring models in improving cash flow and collections. Bus. Credit. **100**, 16–17 (1998)
5. Pozzolo, A.D., Caelen, O., Borgne, Y.L., Waterschoot, S., Bontempi, G.: Learned lessons in credit card fraud detection from a practitioner perspective. Expert Syst. Appl. **41**, 4915–4928 (2014)
6. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. **6**, 20–29 (2004)
7. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**, 429–449 (2002)
8. Lessmann, S., Baesens, B., Seow, H., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. Eur. J. Oper. Res. **247**, 124–136 (2015)
9. Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. **39**, 3446–3453 (2012)
10. Bhattacharyya, S., Jha, S., Tharakunnel, K.K., Westland, J.C.: Data mining for credit card fraud: a comparative study. Decis. Support. Syst. **50**, 602–613 (2011)
11. Saia, R., Carta, S.: A linear-dependence-based approach to design proactive credit scoring models. In: Fred, A.L.N., Dietz, J.L.G., Aveiro, D., Liu, K., Bernardino, J., Filipe, J. (eds.) Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), KDIR, vol. 1, Porto, Portugal, 9–11 November 2016, pp. 111–120. SciTePress (2016)
12. Doumpos, M., Zopounidis, C.: Credit scoring. In: Doumpos, M., Zopounidis, C. (eds.) Multicriteria Analysis in Finance, pp. 43–59. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05864-1_4

13. Ali, S., Smith, K.A.: On learning algorithm selection for classification. Appl. Soft Comput. **6**, 119–138 (2006)
14. Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. **77**, 103–123 (2009)
15. Siami, M., Hajimohammadi, Z., et al.: Credit scoring in banks and financial institutions via data mining techniques: a literature review. J. AI Data Min. **1**, 119–129 (2013)
16. Chen, S.Y., Liu, X.: The contribution of data mining to information science. J. Inf. Sci. **30**, 550–558 (2004)
17. Alborzi, M., Khanbabaei, M.: Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method. IJBIS **23**, 1–22 (2016)
18. Reichert, A.K., Cho, C.C., Wagner, G.M.: An examination of the conceptual issues involved in developing credit-scoring models. J. Bus. Econ. Stat. **1**, 101–114 (1983)
19. Henley, W.E.: Statistical aspects of credit scoring. Ph.D. thesis, Open University (1994)
20. Desai, V.S., Crook, J.N., Overstreet, G.A.: A comparison of neural networks and linear scoring models in the credit union environment. Eur. J. Oper. Res. **95**, 24–37 (1996)
21. Blanco-Oliver, A., Pino-Mejías, R., Lara-Rubio, J., Rayo, S.: Credit scoring models for the microfinance industry using neural networks: evidence from Peru. Expert Syst. Appl. **40**, 356–364 (2013)
22. Henley, W.: A k-nearest-neighbour classifier for assessing consumer credit risk. Statistician **45**, 77–95 (1996)
23. Ong, C.S., Huang, J.J., Tzeng, G.H.: Building credit scoring models using genetic programming. Expert. Syst. Appl. **29**, 41–47 (2005)
24. Chi, B., Hsu, C.: A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. Expert Syst. Appl. **39**, 2650–2661 (2012)
25. Saia, R., Carta, S.: An entropy based algorithm for credit scoring. In: Tjoa, A.M., Xu, L.D., Raffai, M., Novak, N.M. (eds.) CONFENIS 2016. LNBIP, vol. 268, pp. 263–276. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49944-4_20
26. Davis, R., Edelman, D., Gammerman, A.: Machine-learning algorithms for credit-card applications. IMA J. Manag. Math. **4**, 43–51 (1992)
27. Wang, G., Ma, J., Huang, L., Xu, K.: Two credit scoring models based on dual strategy ensemble trees. Knowl.-Based Syst. **26**, 61–68 (2012)
28. Hsieh, N.C.: Hybrid mining approach in the design of credit scoring models. Expert. Syst. Appl. **28**, 655–665 (2005)
29. Lee, T.S., Chen, I.F.: A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. Expert. Syst. Appl. **28**, 743–752 (2005)
30. Wang, G., Hao, J., Ma, J., Jiang, H.: A comparative assessment of ensemble learning for credit scoring. Expert Syst. Appl. **38**, 223–230 (2011)
31. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**, 1263–1284 (2009)
32. Vinciotti, V., Hand, D.J.: Scorecard construction with unbalanced class sizes. J. Iran. Stat. Soc. **2**, 189–205 (2003)
33. Marqués, A.I., García, V., Sánchez, J.S.: On the suitability of resampling techniques for the class imbalance problem in credit scoring. JORS **64**, 1060–1070 (2013)

34. Crone, S.F., Finlay, S.: Instance sampling in credit scoring: an empirical study of sample size and balancing. Int. J. Forecast. **28**, 224–238 (2012)
35. Zhu, J., Wang, H., Yao, T., Tsou, B.K.: Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In: Scott, D., Uszkoreit, H. (eds.) COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18–22 August 2008, Manchester, UK, pp. 1137–1144 (2008)
36. Donmez, P., Carbonell, J.G., Bennett, P.N.: Dual strategy active learning. In: Kok, J.N., Koronacki, J., Mantaras, R.L., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 116–127. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74958-5_14
37. Attenberg, J., Provost, F.J.: Inactive learning? Difficulties employing active learning in practice. SIGKDD Explor. **12**, 36–41 (2010)
38. Thanuja, V., Venkateswarlu, B., Anjaneyulu, G.: Applications of data mining in customer relationship management. J. Comput. Math. Sci. **2**, 399–580 (2011)
39. Lika, B., Kolomvatsos, K., Hadjiefthymiades, S.: Facing the cold start problem in recommender systems. Expert Syst. Appl. **41**, 2065–2073 (2014)
40. Son, L.H.: Dealing with the new user cold-start problem in recommender systems: a comparative review. Inf. Syst. **58**, 87–104 (2016)
41. Fernández-Tobías, I., Tomeo, P., Cantador, I., Noia, T.D., Sciascio, E.D.: Accuracy and diversity in cross-domain recommendations for cold-start users with positive-only feedback. In: Sen, S., Geyer, W., Freyne, J., Castells, P. (eds.) Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, 15–19 September 2016, pp. 119–122. ACM (2016)
42. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
43. Moler, C.B.: Numerical Computing with MATLAB. SIAM, Philadelphia (2004)
44. Quah, J.T.S., Sriganesh, M.: Real-time credit card fraud detection using computational intelligence. Expert Syst. Appl. **35**, 1721–1732 (2008)
45. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**, 107–113 (2008)
46. Shannon, C.E.: A mathematical theory of communication. Mob. Comput. Commun. Rev. **5**, 3–55 (2001)
47. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**, 131–156 (1997)
48. Kwak, N., Choi, C.: Input feature selection for classification problems. IEEE Trans. Neural Netw. **13**, 143–159 (2002)
49. Jiang, F., Sui, Y., Zhou, L.: A relative decision entropy-based feature selection approach. Pattern Recognit. **48**, 2151–2163 (2015)
50. Powers, D.M.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation (2011)
51. Faraggi, D., Reiser, B.: Estimation of the area under the ROC curve. Stat. Med. **21**, 3093–3106 (2002)
52. Salzberg, S.: On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min. Knowl. Discov. **1**, 317–328 (1997)
53. Liu, Y., Schumann, M.: Data mining feature selection for credit scoring models. J. Oper. Res. Soc. **56**, 1099–1108 (2005)

Text Mining for Word Sentiment Detection

Kevin Labille(✉), Susan Gauch, and Sultan Alfarhood

University of Arkansas, Fayetteville, AR 72701, USA
kclabill@uark.edu

Abstract. This work presents a novel approach for automatically generating a sentiment lexicon. We employ an unsupervised learning approach using several probabilistic and information theoretic models. While most of the unsupervised approaches require a set of seed words to begin their work, our methods differ from these by using no *a priori* knowledge. In addition, our models are effective with a diverse corpus rather than requiring a corpus for a limited domain. We demonstrate the effectiveness of our approaches by performing sentiment analysis on Amazon products reviews, comparing the various automatically-generated lexicons. Based on our cross validation results, we show that our lexicons outperform a widely-used sentiment lexicon on both balanced and unbalanced datasets.

Keywords: Sentiment lexicon · Sentiment analysis
Information theory · Text-mining

1 Introduction

Seeking opinions from other people when we need to make decision has long been part of the human experience [23]. You might have asked the following questions to your friends before, or your might have asked them to yourself: Is this item worth buying? Where is the best pizza in town? Until recently, we could only ask those close to us, e.g., neighbors, friends, or family for their thoughts. The rapid growth of online commerce, or e-commerce, has allowed online retailers to make it possible for customers to share their opinions about products and items. One issue is that it is hard to exactly define what an opinion is. The difference between an opinion or a fact is very thin and people will often disagree on what is which [17,18]. Despite this, opinions can be useful not only to online e-commerce but also in government intelligence, business intelligence, and other online services [33] that benefit from summarizing and analyzing collective viewpoints.

The number of online reviews has increased tremendously over the years, and it is now possible to read the opinions of thousands of people all over the Internet on movies, restaurants, hotels, books, products, and professionals. The large amount of information available online today allows researchers to study how individuals express opinions and to mine the collections of opinions to identify

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 149–170, 2019.
https://doi.org/10.1007/978-3-319-99701-8_7

trends and consensus. A new task arose from this phenomenon: sentiment analysis. Sentiment analysis can be divided in two distinct subtasks: opinion summarization and opinion mining. Opinion summarization consists of identifying and extracting products features from user's reviews whereas opinion mining consists of identifying the semantic orientation (positive/negative) of users' reviews.

We traditionally divide sentiment analysis approaches into two categories: corpus-based approaches and the lexicon-based approaches. The first category consists of building classifiers from labeled instances and is often described as a machine-learning approach also known as supervised classification. The latter uses a dictionary of opinion-bearing words, that is, a list of word associated with a sentiment orientation (positive/negative) that is often associated with a sentiment strength as well. Sentiment lexicons are therefore essential to the sentiment analysis task and the accuracy of the resulting sentiment analysis task is dependent upon the quality of the opinion lexicon. If the lexicon is missing words that express sentiment, or if the strength of the sentiments indicated by the words are incorrect, the accuracy of the resulting sentiment analysis will be negatively impacted. High quality sentiment lexicons are now available, however they tend to focus only on adjectives and they store sentiment weights that are applicable to generic opinions. One advantage of these lexicons is that they can be in other where there may not be enough information to do corpus-based approaches.

Our work focuses on generating a sentiment lexicon automatically without *a priori* knowledge from a corpus of documents. Our lexicons have the advantage of being tuned to the subtleties of sentiment expressed in the particular corpus for which they are built. We introduce and evaluate two approaches to perform this task: (1) a probabilistic approach; and (2) an information theoretic approach. We later on combine the best resulting lexicons into a single ensemble lexicon to take advantages of both methods. Our approaches differ from the state-of-the-art in several ways: (a) we generate a lexicon using text mining with no *a priori* knowledge rather than expanding a list of seed words; (b) unlike most of the existing lexicons that contain only adjectives [37], our lexicon includes words from all parts-of-speech; and (c) we use a large diverse corpus rather than a domain-specific corpus.

We evaluate the effectiveness of our methods on balanced and unbalanced datasets through sentiment analysis on Amazon product reviews. Similar to [11], we accumulate the sentiment scores for each word of the review to compute an overall sentiment score. If the score is positive then the review is deemed to be positive; conversely, if the resulting score is negative the review is deemed to be negative. Results show that our methods outperform the baseline approach on both balanced and unbalanced datasets. we achieve an accuracy ranging from 81.36% to 84.60% on a balanced dataset versus 68.37% for the baseline approach and an accuracy ranging from 87.17% to 88.66% on unbalanced dataset against 81.63% for the baseline approach.

The rest of the paper is organized as follow: In Sect. 2, we present various existing work on sentiment analysis and lexicon generation. Section 3 describes the baseline that we use and both our systems (1) and (2). Section 4 contains experimental evaluation and results that we obtain, and Sect. 5 discuss our findings and observations.

2 Literature Review

Mining online opinions and reviews has become a hot research topic in recent years. It is divided into two tasks: opinion summarization and opinion mining. The former consists of identifying and extracting product features from product's reviews in order to summarize them. Hu and Liu [11] proposed a method to find and extract key features and the opinions related to them among several reviews. In contrast, opinion mining consists of analyzing a product's review in order to determine whether or not it reflects a positive or negative sentiment [19,22]. The state-of-the-art distinguishes two ways of performing sentiment analysis, using either supervised learning techniques or unsupervised learning techniques.

Supervised learning techniques is seen as a two-class classification problem and we typically use a naive Bayes classifier or build a Support Vector Machine (SVM) that is trained on a particular dataset [8,21,24,29,31,34,41]. It has been showed that these techniques performs well on the domain for which it is trained. Unsupervised learning techniques consists of computing the semantic orientation of a review from the semantic orientation of each word found in that review. It is referred as a lexicon-based approach since it typically uses sentiment lexicons [1,6,12,15,37,38].

Sentiment rating prediction, or rating-inference, differs from sentiment analysis by focusing on the task of predicting the rating rather than the overall sentiment orientation. Indeed, It is not uncommon to have reviews that are rated within a range, e.g., from 1 to 5, to express a degree of positiveness or negativeness. Pang and Lee [32] tackled this problem using an SVM regression approach and a SVM multiclass approach. Goldberg and Zhu [9] implemented a graph-based semi-supervised approach and improved upon the previous work.

Most of the aforementioned work is done of the document level, that is, they evaluate the sentiment of the entire document (or review), but is it is important to mention sentiment classification on a sentence level i.e., evaluating the sentiment orientation of a single sentence. Both supervised learning and unsupervised learning approaches are suitable for this task. Yu and Hatzivassiloglou [40] used three unsupervised statistical techniques to identify the polarity of a sentence. More recently, Davidov et al. [5] studied the classification of tweets using supervised learning on text, hashtags and smileys.

Another application of sentiment analysis aims to evaluate a particular aspect or feature of a review as opposed to evaluating the sentiment of the whole review. Ding et al. employed a sentiment lexicon in their approach [6] whereas Wei and Gulla [39] modeled the problem as a hierarchical classification problem and utilized a Sentiment Ontology Tree.

Sentiment lexicon can be applied at every level of sentiment analysis, it is therefore important to accurately capture the sentiment of each word in the document, sentence, or review. There are traditionally three ways of generating such a lexicon: (1) manually; (2) using a dictionary; or (3) using a corpus of documents. Dictionary-based approaches typically use a seed word expansion technique. A list of seed words for which the sentiment orientation is known is expanded by searching within a dictionary for the synonyms and antonyms of the seed words. The process is then repeated until the lexicon has grown to a sufficient size [13,28,35]. Corpus-based approaches can also use a list of seed words that is expanded by using a domain corpus rather than a dictionary. Another method consists of adapting a general sentiment lexicon to a domain-specific one by using a domain corpus as well [4,10,14].

Most of the state-of-the-art techniques need some form of *a-priori* knowledge to generate their lexicon. Paltoglou and Thelwall [30] tackled the problem of generating a lexicon without *a-priori* knowledge by using information retrieval weighting schemes to estimate the score of a word. Their work extends the SMART retrieval system and the BM25 probabilistic model by introducing a delta (Δ) variant and smoothed delta variant of the idf. Similarly, Kim et al. [16] tackled this problem by using a term weighting scheme based on corpus statistics as well as contextual and topic related characteristics. A probabilistic approach is used for evaluating the sentiment degree of a document. They evaluate the likelihood of a query given a word using Latent Semantic Analysis (LSA) and Pointwise Mutual Information (PMI). Additionally, they estimate the probability of a document to generate a particular word using the Vector Space (VS) model, the BM25 probabilistic model and Language Modeling (LM) model.

Our method differs from the aforementioned work by (1) introducing a new weighting scheme called brtf.idf and (2) by using Bayes theorem for text classification as our probabilistic approach rather than using the BM25 or LM model. Taking a similar approach, Martineau and Finin [25] introduced Delta tf.idf which basically calculates the difference of a word's tf.idf score in the positive and negative training dataset. Our work extends from this by estimating the score of a word in an unbalanced dataset as rather than requiring a balanced dataset. We also incorporate a parameter to allow us to weight words occurring in more extreme reviews, i.e., 1^* and 5^*, more highly.

3 Generating a Sentiment Score

The sentiment score for a word w is obtained by combining a probabilistic score $Score_{prob}(w)$ and an information theoretic score $Score_{it}(w)$. In the following section, we describe three approaches for each score calculation. The first is based on probability theory, the second on information theory, and the third approach employs an ensemble of sentiment analysers.

3.1 Using Probabilities

The first probabilistic method is based on Baye's theorem [3] that calculates the posterior probability, defined as the probability of an event A happening given that event B has happened. The probabilistic score, $Score_{prob}(w)$, of a word w is introduced by Labille et al. [20]. It is the difference between its probability of being positive, $p(pos|w)$, and its probability of being negative, $p(neg|w)$, as follows:

$$Score_{prob}(w) = p(pos|w) - p(neg|w)$$

where:

$$p(pos|w) = \frac{p(pos) \times p(w|pos)}{p(w)}$$

$$p(neg|w) = \frac{p(neg) \times p(w|neg)}{p(w)}$$

$$p(pos) = \sum_{w'} \sum_{r \in R_{pos}} n_{w'r}$$

$$p(neg) = \sum_{w'} \sum_{r \in R_{neg}} n_{w'r}$$

$$p(w) = \sum_{r \in R} n_{wr}$$

$p(pos)$ is the prior probability of the positive class, i.e., the proportion of words in the corpus that belong to the positive class, $p(neg)$ is the proportion of words that belongs to the negative class, and $p(w)$ is the total number of occurrences of w. Furthermore, $p(w|pos)$ is the posterior probability of w given the positive class and $p(w|neg)$ is the posterior probability of w given the negative class. The formula yields scores in the range from -1 to 1, with the range from -1 to 0 indicating that a word is negative while scores in the range 0 to $+1$ indicate that the word is positive.

We propose 3 different ways of calculating the posterior probability of a word w given the positive or negative class. The first is the simplest:

$$p(w|pos) = \frac{p(w_{pos})}{p(pos)} \tag{P1}$$

$$p(w|neg) = \frac{p(w_{neg})}{p(neg)}$$

where $p(w_{pos})$ is number of times word w appears in the positive class, $p(pos)$ is the proportion of words that belong to the positive class, $p(w_{neg})$ is the number of times w appears in the negative class, and $p(neg)$ is the proportion of words that belong to the negative class. We expect this formula to have difficulty accurately working with unbalanced datasets, e.g., datasets such as Amazon reviews that contain many more positive examples than negative ones.

Our second approach is influenced by Frank and Bouckaert [7] who studied problems arising from using Baye's theorem for text classification with unbalanced classes and proposed a solution. Based on their work, the second method estimates the probability of word w to be positive or negative as follows:

$$p(w|pos) = \frac{\dfrac{\sum\limits_{r \in R_{pos}} n_{wr}}{\sum\limits_{w'} \sum\limits_{r \in R_{pos}} n_{w'r}} + 1}{k_{pos} + 1} \tag{P2}$$

$$p(w|neg) = \frac{\dfrac{\sum\limits_{r \in R_{neg}} n_{wr}}{\sum\limits_{w'} \sum\limits_{r \in R_{neg}} n_{w'r}} + 1}{k_{neg} + 1}$$

where:

$$\sum_{r \in R_{pos}} n_{wr} = n_{w5^*} + n_{w4^*}$$

$$\sum_{r \in R_{neg}} n_{wr} = n_{w1^*} + n_{w2^*}$$

In this approach, $\sum_{r \in R_{pos}} n_{wr}$ is the number of times word w appears in the positive class (i.e., the number of times it appears in each positive review r in corpus R), $\sum_{r \in R_{neg}} n_{wr}$ is the number of times w appears in the negative class, $\sum_{w'} \sum_{r \in R_{pos}} n_{w'r}$ is the number of occurrences of every word in the positive class, and $\sum_{w'} \sum_{r \in D_{neg}} n_{w'r}$ the number of occurrences of every words in the negative class.

Our third probability-based method computes $p(w|pos)$ and $p(w|neg)$ similarly to (2). The only difference is that we add a weight factor γ to take into account the frequency of the words within the 1^* and 5^* review classes. Our intuition is that, since 1^* reviews are more negative than 2^* reviews and 5^* are more positive than 4^* reviews, word occurrences in these more extreme reviews should count for more. Thus, $\sum_{r \in R_{pos}} n_{wr}$ and $\sum_{r \in R_{neg}} n_{wr}$ in our third method become:

$$\sum_{r \in R_{pos}} n_{wr} = \gamma \; n_{w5^*} + n_{w4^*}$$

$$\sum_{r \in R_{neg}} n_{wr} = \gamma \; n_{w1^*} + n_{w2^*}$$

and $p(w|pos)$ and $p(w|neg)$ become:

$$p(w|pos) = \frac{\dfrac{\sum\limits_{r \in R_{pos}} n_{wr}}{\sum\limits_{w'} \sum\limits_{r \in R_{pos}} n_{w'r}} + 1}{k_{pos} + 1} \tag{P3}$$

$$p(w|neg) = \frac{\dfrac{\sum\limits_{r \in R_{neg}} n_{wr}}{\sum\limits_{w'} \sum\limits_{r \in R_{neg}} n_{w'r}} + 1}{k_{neg} + 1}$$

3.2 Using Information Theory

The information theoretic formulae are based on a traditional information theoretic formula called TF-IDF (Term Frequency-Inverse Document Frequency) [36] that assesses the importance of a word when representing the content of a document. As before, the score of a word w is defined as the difference between its positive score and its negative score. Labille et al. [20] introduced the formula as follows:

$$Score_{IT}(w) = \Big(pos(w) - neg(w)\Big) \times IDF(w)$$

$$where:$$

$$IDF(w) = \log \frac{N}{df_w}$$

Once again, we propose 3 formulae to compute the positive and negative score of word w, this time based on variations of TF-IDF. The first uses the traditional relative term frequency of a word and is inspired by [25]

$$\begin{cases} pos(w) & = rtf(w_{5^*}) + rtf(w_{4^*}) \\ neg(w) & = rtf(w_{1^*}) + rtf(w_{2^*}) \end{cases} \tag{I1}$$

where:

$$rtf(w_{x^*}) = \sum_{r_x \in R} \frac{n_{wr}}{|r|}$$

Here, $rtf(wc)$ is the relative term frequency of word w in class c where $c \in 1^*, 2^*, 4^* or, 5^*$; N_{neg} is the total number of negative review; N_{pos} is the total number of positive reviews; N is the total number of reviews. For example, $rtf(w_{5^*})$ is the relative term frequency of w in the 5-star class; and $|r|$ is the size of the review.

As in the case with our initial probability formula, P1, this formula does not account for an unbalanced dataset.

Our second information-theoretic formula adapts to unbalanced data sets. We first introduce a new term called *balanced relative term frequency* or *brtf*, of a word that is a modified relative term frequency that takes into account the unbalanced factor of a word in the dataset. *balanced relative term frequency* computes a word's frequency relative to the type of review it is, that is, a positive or negative review. If a word w belongs to a negative review the brtf is defined as follows:

$$brtf(w_c) = \frac{rtf_{wr}}{N_{neg}} \times N$$

Conversely, if w belongs to a positive review the brtf of w becomes the following:

$$brtf(w_c) = \frac{rtf_{wr}}{N_{pos}} \times N$$

The positive score, $pos(w)$, and negative score, $neg(w)$ of a word become:

$$\begin{cases} pos(w) & = brtf(w_{5^*}) + brtf(w_{4^*}) \\ neg(w) & = brtf(w_{1^*}) + brtf(w_{2^*}) \end{cases} \tag{I2}$$

Finally, based on the same intuition as with the probabilistic approaches, we add a weight factor γ to take into account the frequency of the words within the more extreme review classes 1^* and 5^*. In this case, the positive score and negative scores of a word are now calculated as follows:

$$\begin{cases} pos(w) & = \gamma \ brtf_c(w_{5^*}) + brtf_c(w_{4^*}) \\ neg(w) & = \gamma \ brtf_c(w_{1^*}) + brtf_c(w_{2^*}) \end{cases} \tag{I3}$$

3.3 Ensemble

Since our probabilistic approach uses the global frequency of a word, it gives importance to the distribution of that word on a corpus level while our information theoretic approach incorporates statistics from the importance of words at the document level.

In order to benefit from both methods, we combine the best probabilistic approach with the best information theoretic approach into what we call an ensemble approach. The score of a word w is calculated as the average of both the $Score_{prob}$ and $Score_{IT}$ of that word. Thus, the final score of a word w is calculated as follow:

$$Score(w) = \frac{Score_{prob}(w) + Score_{IT}(w)}{2}$$

4 Evaluation Through Sentiment Analysis

4.1 Experimental Setup and Dataset

Since our goal is to produce a broadly applicable sentiment lexicon, we build them from a large and diverse dataset. We construct our lexicons from Amazon product reviews [26, 27] for 15 different categories to ensure the heterogeneity of the data. We merge reviews ranging from January 2013 through July 2014 from each of the 15 categories into two large datasets: a balanced dataset and an unbalanced dataset. In the balanced dataset we merged the equal numbers of positive and negative reviews whereas in the unbalanced datasets we merged reviews in the same ratio in which they occurred in the dataset that skews heavily positive. We split both datasets into two subsets using 80% for training and the remaining 20% for test purposes. The balanced dataset contains 2,656,872 reviews which are rated from 1 to 5 while the unbalanced dataset contains 11,129,382 reviews. Tables 1 and 2 [20] present some statistics about both datasets.

Table 1. Balanced training and test dataset statistics.

	Training dataset	Test dataset
Number of reviews	2,125,497	531,375
Number of negative reviews	1,062,972	265,464
Number of positive reviews	1,062,525	265,911
Number of 1* reviews	622,915	155,743
Number of 2* reviews	440,057	109,721
Number of 4* reviews	254,294	63,328
Number of 5* reviews	808,231	202,583

We consider 4-star and 5-star reviews to be positive, conversely, 1-star and 2-star reviews are considered negative. We consider that 3-star reviews are neither positive nor negative and are therefore ignored during any experiments.

We evaluate the effectiveness of our lexicons on both the balanced test dataset (531,375 reviews) and unbalanced test dataset (2,225,877 reviews) using a basic sentiment analysis method. The overall score of a review is computed by summing up each word score in the lexicon and by then dividing by the number of words in the review, to normalize for length. If the resulting score is positive, then the review is deemed to be positive; conversely, if the score is negative the review is deemed to be negative.

We compare our results against a baseline lexicon derived from the widely used lexical resource, SentiWorNet [2]. SentiWordNet is constructed using state-of-the-art techniques and it assigns two sentiment scores (a positive score and a negative score) to each word whilst our sentiment lexicon only assigns one score. To account for that, each SentiWordNet word's score is averaged using

Table 2. Unbalanced training and test dataset statistics.

	Training dataset	Test dataset
Number of reviews	8,903,505	2,225,877
Number of negative reviews	1,062,522	265,914
Number of positive reviews	7,063,481	1,766,132
Number of 1* reviews	622,970	155,688
Number of 2* reviews	439,552	110,226
Number of 4* reviews	1,693,861	422,946
Number of 5* reviews	5,369,620	1,343,186

Petter Tonberg's sentiment value approximation (source code available on SentiWordNet's website). Finally, SentiWordNet takes the POS tag of each word into consideration while our lexicons do not. To account for that, each word's score is averaged across all POS tag which result in a single sentiment score.

4.2 Experimental Results

Balanced Dataset

We first investigate the impact of the γ factor on the accuracy of our ensemble formulae. We measure the accuracy in every classes, i.e., 1-star, 2-star, 4-star, and 5-star, and compare them to the overall accuracy of the system. Figure 1, adapted from [20], depicts the various accuracy measurements for different values of γ. We can notice that the accuracy of the negative class (1-star and 2-star) decreases as γ increases. Conversely, the accuracy of the positive class increases as γ increases. We also notice that when γ is very low, i.e., equal to 0.5, the system is highly accurate in the negative class (97% accuracy for the 1-star class

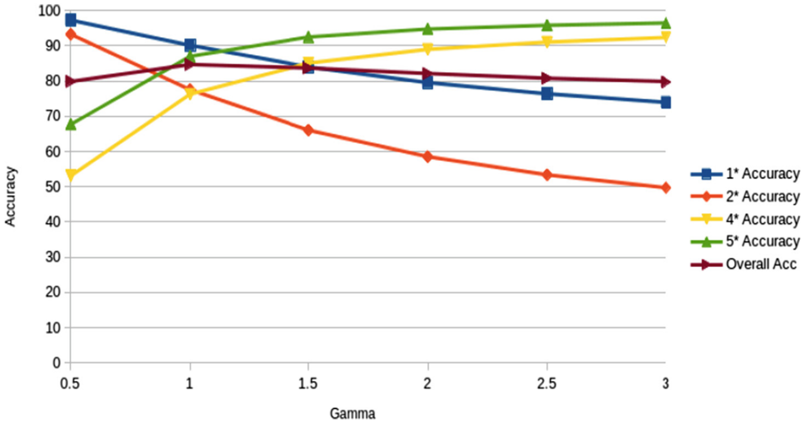


Fig. 1. Impact of Gamma on the accuracy for a balanced dataseti adapted from [20].

and 93% for the 2-star class) and highly inaccurate in the positive class (53% accurate for the 4-star class and 67% accurate for the 5-star class) with an overall accuracy of 79.88%. Conversely, if γ is too high, i.e., equal to 3, the system is highly accurate in the positive class (92% in the 4-star class and 96% in the 5-star class) and poorly accurate in the negative class (74% in the 1-star class and 50% in the 2-star class) with an overall accuracy of 79.70%.

We can observe a balanced accuracy between the positive class and negative class when γ is equal to 1, that is, when the extremes classes (1-star and 5-star) are weighted equally to the middle classes (2-star and 4-star).

When γ is set to 1, not only the average accuracy of the negative class (83%) is balanced with the average accuracy of the positive class (82%), but we also achieve our highest overall accuracy of 84.66%. These observations suggest that we do not need to distinguish between the subclasses when the dataset is balanced.

Table 3. Comparison of the different formulae on balanced dataset.

	TPR	TNR	PPV	NPV	F-Score	Accuracy
P1	0.65	0.95	0.92	0.73	0.76	80.14%
P2	0.84	0.84	0.84	0.84	**0.84**	**84.38%**
P3 ($\gamma = 1$)	0.84	0.84	0.84	0.84	**0.84**	**84.38%**
I1	0.79	0.82	0.81	0.80	**0.81**	**81.36%**
I2	0.80	0.82	0.82	0.80	**0.81**	**81.36%**
I3 ($\gamma = 1$)	0.80	0.82	0.81	0.80	**0.81**	**81.36%**

Table 3 presents the evaluation and comparison of the different formulae on the balanced dataset. We report the True Positive Rate (TPR) that measures the proportion of positive reviews that are correctly classified as positive, the True Negative Rate (TNR) that measures the proportion of negative reviews that are properly classified as negative. We report the Predicted Positive Value (PPV) that measures the proportion of positive results that are true positive and the Negative Predictive Value (NPV) that measures the proportion of negative results that are true negative. We also report the F1-Score and the accuracy. All of the reported values are the averaged values based on a 5-fold cross validation. Since γ is equal to 1, I2 is the same as I3 and P2 is the same as P3, i.e., the extreme reviews are not weighted more heavily.

As shown in Table 3, all formulae achieve a high accuracy on the balanced dataset. P1 and I1 achieve an accuracy of 80.14% and 81.36% respectively, confirming our intuition that they work well on a balanced dataset. P1 has a high TNR and a fairly low TPR whilst I1's TPR and TNR are comparatively similar, suggesting that even though the probabilistic formula performs very well on classifying negative reviews, the information theoretic formula is more reliable for classification. We further notice that P2 and I2 both outperform P1 and I1 in

Table 4. Comparison of the different approaches on balanced dataset.

	Recall	Precision	F1-Score	Accuracy
P3 ($\gamma = 1$)	0.84	0.84	**0.84**	84.38%
I3 ($\gamma = 1$)	0.80	0.82	0.81	81.36%
Ensemble (I3 + P3)	0.83	**0.85**	**0.84**	**84.60%**
Baseline	**0.86**	0.63	0.73	68.37%

all metrics, meaning that our enhanced probabilistic and information theoretic formulae, although designed for unbalanced datasets, are also more accurate on balanced datasets.

We then compare our different individual approaches to the ensemble approach and our baseline. The ensemble approach is built by combining both best individual approaches: I3 and P3 with $\gamma = 1$. Table 4 sums up the comparison using the recall, precision, F1-Score and accuracy.

As we can see, all of our approaches outperform the baseline approach, with the ensemble approach being the most accurate with an accuracy of 84.60% and a F1-Score of 0.84 compared to an accuracy of 68.37% and a F1-Score of 0.73 the baseline, improving the accuracy by 16.23%.

Unbalanced Dataset

We now explore the effectiveness of our approaches on unbalanced datasets. As in the previous section, we first look at the impact of γ on the positive accuracy, negative accuracy and the overall accuracy of the system. Because we are now using the unbalanced dataset, we expected γ to have an impact on the effectiveness of our formulae.

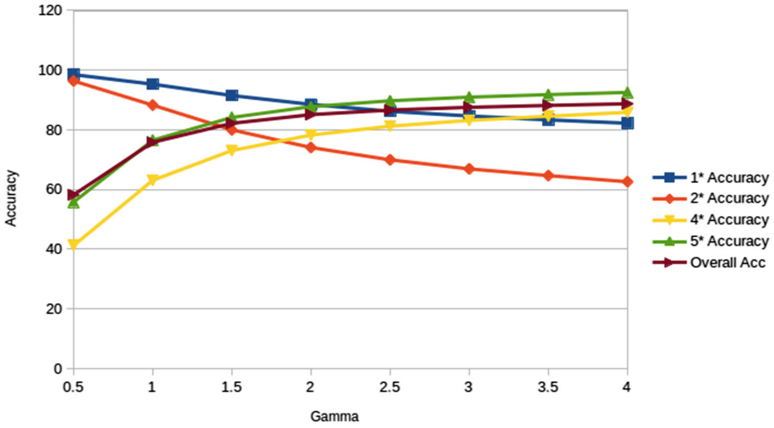


Fig. 2. Impact of Gamma on the accuracy for an unbalanced dataset from [20].

Figure 2, obtained from [20], shows the effect of γ on the accuracy of our ensemble approach on unbalanced dataset. Similarly to the balanced dataset, when γ increases the system becomes more accurate in the positive class and a bit less accurate in the negative class. When γ is at its lowest, i.e., 0.5, the system is extremely accurate in the negative class (97%) but not accurate at all in the positive class (48%) with an overall accuracy of 58.25%.

As γ increases, the system becomes generally more accurate. We achieve our best accuracy when γ is set to 4, with an overall accuracy of 88.75%, an accuracy in the positive class of 89%, and an accuracy in the negative class of 72%.

We can further notice that when γ is set to 1, that is when P2 = I2 and P3 = I3, our system achieves 75.84% accuracy which is less than in the balanced dataset, showing the importance of the γ factor when dealing with unbalanced datasets.

Table 5. Comparison of the different formulae on unbalanced dataset.

	TPR	TNR	PPV	NPV	F1-Score	Accuracy
P1	1.0	0.0	0.86	0.0	**0.92**	86.92%
P2	0.69	0.94	0.98	0.31	0.81	73.30%
P3 ($\gamma = 4$)	0.88	0.79	0.96	0.50	**0.92**	**87.17%**
I1	1.0	0.0	0.86	0.0	**0.92**	86.92%
I2	0.74	0.86	0.97	0.33	0.84	75.96%
I3 ($\gamma = 4$)	0.90	0.67	0.94	0.51	**0.92**	**87.33%**

Table 5 presents the comparison of our formulae on the unbalanced dataset for several metrics (TPR, TNR, PPV, MPV, F1-Score, and accuracy). Each result is the average resulting from a 5-fold cross validation.

Although P1 and I1 both achieve a high accuracy of 86.92%, they have a TPR and NPV of 0.0, meaning that these two approaches are not able to correctly identify negative reviews. They are therefore not suitable for unbalanced datasets. Introducing factors to accommodate for unbalanced dataset in the formulae P2 and I2 allows us to better identify negative reviews as evidenced by the non-null TNR of both P2 and I2. Their high TNR is however offset by a decreased TPR that results in a noticeable loss in accuracy of more than 10% relative to P1 and I1.

By introducing the γ factor in P2 and I2, we are able to increase our TPR and therefore overcome the drop of accuracy. We achieve our highest F1-Score and accuracy of 87.17% for the probabilistic approach and 87.33% for the information theoretic approach while still being able to correctly classify both the negative and positive class.

We compare our approaches to the baseline and report the recall, precision, F1-Score and accuracy in Table 6. Since P3 and I3 are our best probabilistic and information theoretic approaches, they are used for the ensemble approach.

Table 6. Comparison of the different approaches on unbalanced dataset.

	Recall	Precision	F1-Score	Accuracy
P3	0.88	**0.96**	0.92	87.17%
I3	**0.90**	0.94	0.92	87.33%
Ensemble (P3 + I3)	**0.90**	0.95	**0.93**	**88.66%**
Baseline	0.86	0.92	0.89	81.63%

We can notice that all of our approaches outperform the baseline in terms of all metrics used. As with the balanced dataset, our ensemble approach achieves the best accuracy of 88.66% which is an improvement of 7.03% over the baseline that achieves 81.63%.

Finally, we see that the ensemble approach achieves a better recall and precision than the baseline, suggesting that the resulting lexicon could be more exact and complete than the baseline lexicon. The ensemble approach inherits the strength of both the probabilistic approach and information theoretic approach, making it better than each of them individually.

5 Discussion

To give an intuitive feel for the lexicon produced, Table 7 [20] reports the top 5 most positive and most negative words from our lexicons as well as from the baseline lexicon. One very interesting observation is the difference in vocabulary of our lexicons from one approach to another. Indeed, the top words in the probabilistic approach are much more formal than the top words from the information theoretic approach. Likewise, the top words in the baseline approach are uncommon words.

Table 8 [20] shows various words and their relative score across all the lexicons. It is important to note that some words such as *good* are classified as positive in every lexicon. On the other hand, the word *okay* has a negative connotation in our text mining approaches whereas it has a positive connotation in the baseline lexicon. Similarly, *refund* is perceived as negative by all of our approaches while it is perceived as positive in the baseline lexicon. This illustrates the ability of a text mining approach to generate sentiment weights tuned to the dataset, in this case, products reviews.

Table 9 shows some selected words that we classify as **stable** or **variable**. **Stable** words are words that have a constant score, or weight, across lexicons whereas **variable** words are words that have very different or opposite weight across different lexicons. For instance, the words *pretty* and *adequate* are both positive and have a weight of 0.041 in both the baseline lexicon and the ensemble lexicon. Similarly, the words *flimsy* and *inquiry* are both negative words with a weight of −0.034.

We observe two types of **variable** words: (a) words that have the same sentiment orientation but with different strengths; and (b) words that have opposite

Table 7. Top 5 words for each lexicon.

Approach	Top 5 positive words	Top 5 negative words
P3	Perfectible	Garbaged
	Marvellously	Junkiest
	Oustanding	Refundable
	Lushness	Misadvertised
	Grogginess	Defectively
I3	Great	Not
	Love	Waste
	Easy	Money
	Perfect	Return
	Well	Disappointed
Ensemble	Great	Waste
	Love	Money
	Easy	Not
	Perfect	Refund
	Loved	Return
Baseline	Wonderfulness	Angriness
	Fantabulous	Henpecked
	Congratulations	Lamentable
	Excellent	Motormouth
	Bliss	Shitwork

Table 8. Selected words and their score.

	Good	Refund	Okay	Speaker
P3	0.0524	−0.7050	−0.0421	0.0079
I3	0.4384	−0.2422	−0.0421	−0.0075
Ensemble	0.2454	−0.4736	−0.0619	0.0000
Baseline	0.4779	0.0000	0.2500	0.0000

sentiment orientation. Words such as *bliss* or *unsupported* are of type (a) where the former is highly positive in the baseline lexicon and slightly positive in our lexicon whereas the former is highly negative in the baseline lexicon and fairly negative in our lexicon. Words such as *reputable* or *joking* are of type (b), that is, positive in one lexicon and negative in another lexicon.

The example in Table 10 shows the results of sentiment analysis for a review using our various lexicons as well as the baseline lexicon. The review is first preprocessed so as to remove stopwords and any punctuation marks. We then query our lexicon to find each word's score and sum them up to a single score.

Table 9. Sample stable words vs variable words.

Word	Baseline weight	Ensemble weight
Stable words		
Inquiry	−0.034	−0.034
Pretty	0.041	0.041
Adequate	0.011	0.011
Flimsy	−0.170	−0.169
Variable words		
Reputable	0.875	−0.163
Bliss	1.000	0.044
Joking	−0.030	0.875
Unsupported	−0.667	−0.159

Table 10. Comparison of lexicons on rated review.

Review								Overall rating	
Full review	I purchased this based on some of the other reviews but this was crappy and only worked for two days.							**1* negative**	
Review after preprocessing	Purchased	Based	Reviews	Crappy	Worked	Two	Days	Total Score	Classified as
P3 score	−0.02	−0.01	−0.09	−0.29	−0.06	−0.007	−0.03	−0.0764	Neg
Coverage	100%								
I3 score	−0.01	−0.007	−0.12	−0.03	−0.10	−0.05	−0.07	−0.0601	Neg
Coverage	100%								
ensemble score	−0.02	−0.009	−0.11	−0.16	−0.08	−0.03	−0.05	−0.0683	Neg
Coverage	100%								
baseline score	N/A	0.0	N/A	−0.75	N/A	0.0	0.0	−0.1875	Neg
Coverage	57.14%								

The resulting score is then averaged by the number of words present in the lexicon to account for the length of the review. If the score is positive then the review is deemed to be positive, conversely, if the score is negative the review is deemed to be negative.

We also report the coverage, that is, the proportion of words from the review that are found in the lexicon. As shown in the table, our lexicons all have a coverage of 100% against 57% for the baseline. Words that are not covered by the lexicon are noted as N/A. While our approach can score every word in the review, the baseline lexicon misses several words among which could be words carrying important information such as *worked*. This major difference may explain the higher accuracy achieved by our lexicons.

The baseline lexicon focuses on adjectives whereas our approach works on all parts-of speech, based on the belief that non-adjectives can indicate sentiment orientation and that it is important to take them into consideration.

Table 11 shows snippets of reviews properly classified by our best approach, i.e., the ensemble approach, that were misclassified by the baseline lexicon.

Review A is a 2-star review that is classified as negative by our lexicon with a overall score of −0.0608 and classified as positive by the baseline lexicon with a score of 0.0082. It is important to note that our approach has a coverage of 100%, i.e., every words from the review are scored, while the baseline only covers 54.5% of the review's vocabulary. We believe that words such as *worked* or *died* are indicators of negativity in the review and therefore play a role in the sentiment classification of the review. By failing to take them into consideration, the baseline might have failed to correctly classify the review.

Likewise, review B is a 5-star review that was properly classified by our ensemble approach and misclassified by the baseline lexicon. Here again, the baseline lexicon has a low coverage, i.e., 58.3% against 91.66% for our approach. This review reveals another interesting point, that is, the word *unusing* which is a misspelled word by the product's reviewer. Both lexicons fail to score the word, highlighting the importance of spelling and orthography in text mining.

Also note the difference between individual word's scores. For instance, the baseline lexicon evaluates *described* and *quickly* as neutral words whereas they are evaluated as positive in our lexicon. This could explain the misclassification of the review by the baseline lexicon.

Table 12 shows snippets of reviews that are misclassified by both our approaches and the baseline (review C) as well as a review (review D) that is misclassified by our approach but correctly classified by the baseline. Although both our ensemble lexicon and the baseline lexicon have a very high coverage in review C, they still fail to properly classify the review. This is mainly due to the nature of the product review, which is very short, that does not provide enough information to allow a correct classification.

Table 11. Snippets of properly classified reviews.

	Review A												Overall rating	
Full review	I bought this item and it worked for several months and then just died. I had to replace it and didn't have time to reorder.												**2* negative**	
Review after preprocessing	bought	item	worked	several	months	just	died	replace	didnt	time	reorder		Total Score	Classified as
baseline score	N/A	0	N/A	-0.18	N/A	0.33	N/A	-0.24	N/A	0.06	0.06		0.0082	Pos
Coverage	54.50%													
ensemble score	0.01	-0.18	-0.08	0.02	-0.11	0.02	-0.08	-0.03	-0.21	0.00	-0.03		**-0.0608**	**Neg**
Coverage	100%													
	Review B												Overall rating	
Full review	Fits my Fellows Shredder great. Product was as described and delivered quickly. A huge step up from the bags I had been unsing												**5* positive**	
Review after preprocessing	fits	fellows	shredder	great	product	described	delivered	quickly	huge	step	bags	unsing	Total Score	Classified as
baseline score	N/A	N/A	0	0.12	0	0	0	0	-0.125	-0.03	N/A	N/A	-0.0091	Neg
Coverage	58.3%													
ensemble score	0.18	0.02	0.00	0.55	-0.11	0.05	0.02	0.08	-0.00	0.04	0.02	0	**0.08**	**Pos**
Coverage	91.66%													

Indeed, by taking a closer look at review C, we can see that only very few words express a negative thought, that is "wouldn't recommend". From a computing perspective that is only 2 words out of 11 that really carry negativity. Hence the inability for both approaches to perform a correct classification.

Table 12. Snippets of misclassified reviews.

Review C													Overall rating	
Full review	This book has really basic information. I'm pretty sure I could have found this stuff online. I wouldn't recommend it.												1* **negative**	
Review after preprocessing	book	really	basic	information	pretty	sure	found	stuff	online	wouldnt	recommend		Total Score	Classified as
baseline score	0.03	0.47	0.03	0.02	0.04	0.12	0	-0.001	0	N/A	0.13		0.0858	Pos
Coverage	90.90%													
ensemble score	0.27	0.14	0.004	0.02	0.04	0.005	0.03	-0.03	0.11	-0.15	-0.03		0.0562	Pos
Coverage	100%													
Review D													Overall rating	
Full review	This turned out to be a lot smaller than I pictured it and because of the added art it's designed in such a way that only half the capacity can be used.												2* **negative**	
Review after preprocessing	turned	lot	smaller	pictured	added	art	designed	way	half	capacity	can	used	Total Score	Classified as
baseline score	-0.08	-0.01	-0.125	0	N/A	0.12	0	0.002	-0.03	0.03	0	0.07	**-0.001**	**Neg**
Coverage	91.66%													
ensemble score	-0.04	0.10	0.004	-0.06	0.05	0.06	-0.01	0.00	-0.08	0.02	0.15	0.002	0.015	Pos
Coverage	100%													

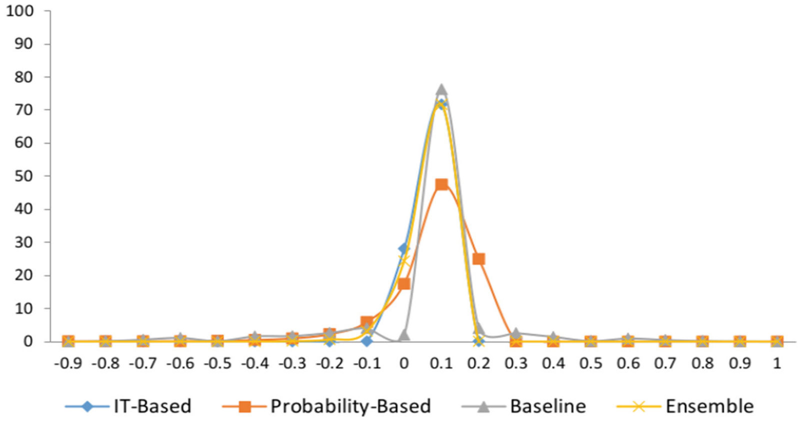


Fig. 3. Score distribution from the different approaches.

Review D shows that our approach can sometimes fail to properly classify a review when the baseline lexicon can. In particular, Review D is a 2-star review, and these tend to be harder to classify since they are on the edge of being neutral.

Figure 3 [20] shows the word sentiment score distribution for each of the approaches. As we can see, all lexicons words score tend to fail in the range -0.25 to 0.25. there are also many words with positive scores versus those with negative scores.

6 Conclusions and Future Work

In conclusion, we describe a new text mining-based technique based on probabilities and information theory to automatically generate a sentiment lexicon. Unlike most state-of-the-art techniques that either use a list of seed words or that perform lexicon adaptation, our method does not require any *avpriori* knowledge. Our approach differs from the traditional techniques in several ways: (1) We use a large diverse corpus to train our model rather than using domain-specific corpus; (2) our lexicons include words from all part-of-speech (POS) rather than being limited to adjectives; and (3) our model is accurate on both balanced datasets and unbalanced datasets.

Our approaches are validated by using our lexicons to run sentiment analysis on Amazon product reviews. Our best probabilistic approach and our best information theoretic approach are combined into an ensemble approach that outperforms each of the individual methods and the baseline. We achieve an accuracy ranging from 81.36% to 84.60% on a balanced dataset versus 68.37% for the baseline approach and an accuracy ranging from 87.17% to 88.66% on unbalanced dataset against 81.63% for the baseline approach. Our method also achieves a good recall, precision, and F1-Score, showing that we are able to classify both negative and positive reviews correctly.

Our future work will focus on the exploration of domain-specific sentiment analysis. We will evaluate the effectiveness of our method across several domains. Our intuition is that some words' sentiments are depending upon the context in which they are used, and a single word can therefore have different or opposite sentiment orientation. Another focus will be exploration of sentiment rating prediction rather than sentiment analysis, i.e., how to predict the class of a review from a set of several classes level rather than being limited to a simply positive or negative.

References

1. Abdulla, N.A., Ahmed, N.A., Shehab, M.A., Al-Ayyoub, M., Al-Kabi, M.N., Al-rifai, S.: Towards improving the lexicon-based approach for Arabic sentiment analysis. Int. J. Inf. Technol. Web Eng. (IJITWE) **9**(3), 55–71 (2014)
2. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC, vol. 10, pp. 2200–2204 (2010)
3. Bayes, M., Price, M.: An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. Philos. Trans. (1683–1775) **53**, 370–418 (1763)
4. Choi, Y., Cardie, C.: Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pp. 590–598. Association for Computational Linguistics (2009)
5. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using Twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 241–249. Association for Computational Linguistics (2010)
6. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 231–240. ACM (2008)

7. Frank, E., Bouckaert, R.R.: Naive Bayes for text classification with unbalanced classes. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 503–510. Springer, Heidelberg (2006). https://doi.org/10.1007/11871637_49
8. Gao, D., Wei, F., Li, W., Liu, X., Zhou, M.: Cross-lingual sentiment lexicon learning with bilingual word graph label propagation. Comput. Linguist. **41**, 21–40 (2015)
9. Goldberg, A.B., Zhu, X.: Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, pp. 45–52. Association for Computational Linguistics (2006)
10. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, pp. 174–181. Association for Computational Linguistics (1997)
11. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
12. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: AAAI, vol. 4, pp. 755–760 (2004)
13. Kamps, J., Marx, M., Mokken, R.J., De Rijke, M., et al.: Using WordNet to measure semantic orientations of adjectives (2004)
14. Kanayama, H., Nasukawa, T.: Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 355–363. Association for Computational Linguistics (2006)
15. Khan, A.Z., Atique, M., Thakare, V.: Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Int. J. Electron. Commun. Soft Comput. Sci. Eng. (IJECSCSE), 89 (2015)
16. Kim, J., Li, J.J., Lee, J.H.: Discovering the discriminative views: measuring term weights for sentiment analysis. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, pp. 253–261. Association for Computational Linguistics (2009)
17. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics, p. 1367. Association for Computational Linguistics (2004)
18. Kim, S.M., Hovy, E.: Extracting opinions, opinion holders, and topics expressed in online news media text. In: Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp. 1–8. Association for Computational Linguistics (2006)
19. Kim, S.M., Hovy, E.: Identifying and analyzing judgment opinions. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 200–207. Association for Computational Linguistics (2006)
20. Labille, K., Alfarhood, S., Gauch, S.: Estimating sentiment via probability and information theory. In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, KDIR, vol. 1, pp. 121–129 (2016). https://doi.org/10.5220/0006072101210129

21. Li, T., Zhang, Y., Sindhwani, V.: A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, pp. 244–252. Association for Computational Linguistics (2009)
22. Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, vol. 2, pp. 627–666 (2010)
23. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
24. Liu, F., Wang, D., Li, B., Liu, Y.: Improving blog polarity classification via topic analysis and adaptive methods. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 309–312. Association for Computational Linguistics (2010)
25. Martineau, J., Finin, T.: Delta TFIDF: an improved feature space for sentiment analysis. ICWSM **9**, 106 (2009)
26. McAuley, J., Pandey, R., Leskovec, J.: Inferring networks of substitutable and complementary products. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM (2015)
27. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–52. ACM (2015)
28. Mohammad, S., Dunne, C., Dorr, B.: Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pp. 599–608. Association for Computational Linguistics (2009)
29. Ng, V., Dasgupta, S., Arifin, S.: Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In: Proceedings of the COLING/ACL on Main Conference Poster Sessions, pp. 611–618. Association for Computational Linguistics (2006)
30. Paltoglou, G., Thelwall, M.: A study of information retrieval weighting schemes for sentiment analysis. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1386–1395. Association for Computational Linguistics (2010)
31. 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 (2004)
32. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115–124. Association for Computational Linguistics (2005)
33. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
34. 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, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
35. Peng, W., Park, D.H.: Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. Urbana **51**, 61801 (2004)

36. Salton, G., McGill, M.J.: Introduction to modern information retrieval (1986)
37. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
38. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics (2002)
39. Wei, W., Gulla, J.A.: Sentiment learning on product reviews via sentiment ontology tree. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 404–413. Association for Computational Linguistics (2010)
40. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 129–136. Association for Computational Linguistics (2003)
41. Zhou, S., Chen, Q., Wang, X.: Active deep networks for semi-supervised sentiment classification. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 1515–1523. Association for Computational Linguistics (2010)

Knowledge Engineering and Ontology Development

A Smart System for Haptic Quality Control: A Knowledge-Based Approach to Formalize the Sense of Touch

Bruno Albert[1,2,4](✉), François De Bertrand De Beuvron[1], Cecilia Zanni-Merk[3], Jean-Luc Maire[2], Maurice Pillet[2], Julien Charrier[4], and Christophe Knecht[4]

[1] ICube Laboratory SDC Team, INSA de Strasbourg,
300 bd Sébastien Brant, 67400 Illkirch, France
[2] SYMME Laboratory, Université Savoie Mont-Blanc,
7 Chemin de Bellevue, 74940 Annecy-Le-Vieux, France
[3] LITIS Laboratory, INSA de Rouen,
Avenue de l'université, 76800 Saint Etienne du Rouvray, France
[4] INEVA, 14 rue du Girlenhirsch, 67400 Illkirch, France
bruno.albert@ineva.fr

Abstract. The field of quality control has seen over the last decades a variety of studies and innovations turned towards the improvement of the perceptions rendered through manufactured products. Thus, quality checks do not only rely on technical control, but on a diversity of controls which correspond to the senses involved when interacting with a product. However, the quality specifications and in particular the vocabulary used for their description are still very specific to each product or industrial domain. With the perspective of simplifying and standardizing perceived quality control, this study aims at providing a Smart System based on knowledge modelling methods which is capable of guiding manufacturers in the process of structuring, generalizing and eventually automatizing the control process related to perceived quality and touch in particular. This paper presents a general framework for the Smart System as well as an ontological structure for the representation of perceived quality knowledge. The specificities of the sense of touch are detailed and led to the proposition of novel formalized description and conceptual model of haptic perceptions.

Keywords: Sensory perception · Haptics · Smart system
Semantic analysis · Quality control · Perceived quality

1 Introduction

Humans perceive the world and particularly objects in the world through their senses. It allows us not only to understand, but also to make our own opinion and judgment about these objects. Perceived quality has become a major factor of the choice of products by customers, and a great economical challenge for

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 173–190, 2019.
https://doi.org/10.1007/978-3-319-99701-8_8

manufacturers. In an industrial context, controlling perceived quality is often limited to controlling the visual quality of products. However, in most cases only controlling visual aspects does not fully correspond to the perception a customer can have when interacting with the product. In particular, the action of touching usually comes right after a first visual observation, and has an important role in completing the full perception [1].

Touch involves complex physical and psychological phenomena which lead to very precise but also very subjective and individual haptic perceptions. Haptic perceptions are a combination of tactile and kinesthetic perceptions [2]. Tactile sensations are obtained thanks to sensory receptors localized in the skin. Kinesthetic sensations are obtained thanks to receptors localized in muscles, tendons and joints. When touching a product, as a simplification one can consider tactile sensations as the sensations obtained locally on the surface of the product and kinesthetic sensations as the ones obtained more globally over the product.

Therefore, on the one hand the control of haptic sensations involves the comprehension of human haptic perception process, at physical as well as psychological levels. On the other hand, it involves the formalization of this knowledge in order to extract control protocols and make this knowledge usable by an automated system. In this context, knowledge based systems are especially suitable.

For the development of a Smart System for haptic quality control the KREM framework was chosen. It is presented in section two. This framework highlights the use of four main components (knowledge, rules, experience, metaknowledge) to take into account the specificities of the system. In particular, this paper details the Knowledge component. An ontological structure is introduced in section three. It presents the different domains of knowledge involved and the corresponding ontologies, as well as the upper level ontology used to structure the concepts. The proposed formalization of haptic perception knowledge is explained in section four. A domain ontology structuring haptic knowledge is then presented in section five, with the integration of the proposed formalization of the haptic domain.

2 A Framework for the Development of the Smart System

Conventionally, a smart system is composed of a fact base and a rule base, on which various types of reasoning can be made. However, the observation of the drawbacks of this classic architecture led to the selection of different model, rather based on the use of semantic technologies. Therefore, the KREM model [3] is used here as a global framework in order to develop the proposed smart system. This framework has already been successfully applied in several different domains [4,5]. It was especiallycially proven to be useful regarding the management of knowledge in applications where the elicitation of expert knowledge and the non-completeness of this elicitation could be a problem [6].

Semantic technologies use methods from automatic language processing, machine learning and knowledge representation to build the ontologies and the rules that will enable its implementation. Semantic technologies also intend to create new meaningful relationships, and therefore new knowledge, based on information of different natures and forms. Semantic technologies offer in particular to enriching documents with meta-data or creating specific linguistic or terminological standards. It can eventually facilitate decision making through effective knowledge management.

But decision-making, to be effective, must result from reasoning and analysis of this knowledge. It must also take into account the experience and expertise of decision-makers. The capitalization of experience appeared naturally as a possibility of improvement of the architecture, in the form of specific knowledge structures and reasoning mechanisms. SOEKS (set of experience knowledge structure) [7] and CBR (case based reasoning) [8] are two examples of these structures and reasoning mechanism.

Furthermore, the use of meta-knowledge has become a need, in order to lead the execution of our knowledge-based systems following the application environment. Meta-knowledge is knowledge about the domain knowledge, the rules or the experience. It can be in the form of context, culture or protocols that steer the use of that knowledge. Context is any information that characterizes a situation related to the interaction between human beings, applications and the surrounding environment [9] and is identified as belonging to four types: identity, status, location, time. Context is typically the location, identity and state of people, groups, and computational and physical objects. Time is information that helps to recognize a situation using historical data. The Culture aspect of meta-knowledge intends to reflect the different ways decisions are made in different cultures. Protocols usually elicits the ways the other pieces of knowledge are used to accomplish a task (for example, quality control). Meta-knowledge may also be closely related to experience knowledge.

To take these ideas into account, the KREM model has four interacting components that can be defined by project or application domain. The re-use of components is, of course, encouraged. The KREM components are (Fig. 1):

- The *Knowledge* component contains the domain knowledge to operate, by means of different domain ontologies.
- The *Rules* component allows different types of reasoning (monotone, spatial, temporal, fuzzy, or other) depending on the application.
- The *Experience* component allows the capitalization and re-use of prior knowledge.
- The *Meta-knowledge* component, including knowledge about the other three bricks. This component depends on the problem.

The way the domain knowledge is formalized defines how the rules are expressed. Experience completes the available knowledge and rules. Finally, meta-knowledge directly interacts with the rules and the experience to indicate which rules (coming from experience or from the initial rule set) can be used according to the context of the problem to solve.

A modular architecture, such as KREM, is one of the main framework for large and complex systems. In this framework, each module or component has a

Fig. 1. The KREM architecture with its four interrelated components: knowledge, rules, experience and meta-knowledge [3].

specific functionality providing separation of components. In turn, this enables the system to support re-use or replacement (*i.e.* changes in a single module would not affect the others, permitting the continuous operation of the system).

Formalizing and structuring Knowledge are the first steps of the development of the Smart System. The following sections are hence focused on the Knowledge component, which will eventually be integrated to a larger KREM architecture and therefore coupled with the other modules previously presented.

3 Towards an Ontological Structure for Haptic Quality Control

The use of formal models, such as ontologies, is essential for the development of a smart system. Very few studies have proposed ontologies directly related to the description of human perceptions and quality control. This is why we propose here a novel way to model knowledge related to the description and control of perceived quality. This section hence presents the proposed ontological structure for the measuring of sensory perceptions and its particularization towards haptic quality control.

3.1 Upper-Level Conceptual Model

The construction of a conceptual domain representation requires first to identify the general ontological structure in which the domain ontology can be included. In particular, the use of a high-level ontology is essential in order to have a "skeleton" of the structure, which gives it a coherent and already tested composition. There are multiple upper-level ontologies, and we have chosen to focus on the Semantic Sensor Network (SSN) ontology [10], supported by the W3C[1]. Indeed, it has been identified as particularly relevant considering the context of the study and the opportunities of further development regarding the instrumentation of the control, but also regarding enrichment through experience. Compton [10] introduced the SSN ontology in order to describe sensors and observations. Besides the perspectives of future development of the present study

[1] World Wide Web Consortium (http://www.w3.org/).

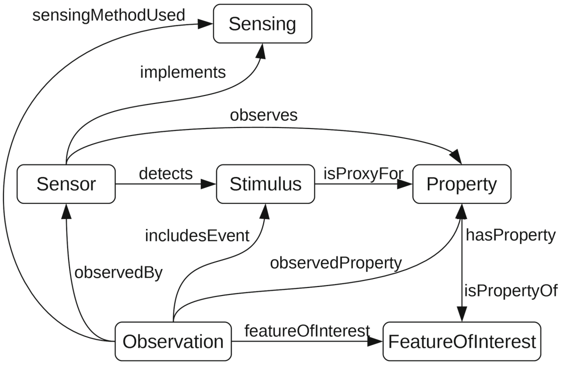


Fig. 2. Extract of the SSN ontology [10].

around the system of sensors, the SSN ontology introduced a way to conceptualize the links between product properties, sensors measurement and observation conditions. In addition, SSN is a core ontology that is based on the well-known top-level DUL ontology [11].

Figure 2 is a reduced version of the SSN ontology including the stimulus-sensor-observation pattern proposed by Compton [10]. This figure focuses only on some of the entities relevant to this study. This representation involves the different concepts of interest, regarding the aim of integrating haptic perception knowledge, as well as future automation of the process. This extract of the SSN ontology and the proposed domain ontology have been aligned. The details of this alignment is presented below.

3.2 Global Ontological Structure

An ontological structure is proposed in Fig. 3, in order to gather and structure the diversity of domains involved in the control of the haptic quality of products. This structure was first introduced in [12] and refined since. It aims at organizing the elements of knowledge that compose each domain into different domain ontologies. They can then be aligned to the upper-level ontology, i.e. the SSN ontology here. This ontological structure is presented as a classical ontology hierarchy [13] with a top-level ontology, a core ontology, a general ontology (which is not detailed here), a task ontology and multiple domains ontologies. The domains involved here are: the application context, the sensing and the sensory perceptions domains (and haptic perceptions in particular). In addition, the control process is represented as a task ontology.

Here is a brief description of all these ontologies:

- The *Application Context* ontology formalizes industrial constraints, product properties and environmental context details.
- The *Sensing* ontology gathers knowledge about sensors, processing methods and possible kinematics which are relevant for the measuring of sensations.

- The *Sensory Perception* ontology gathers knowledge about human senses and enables a direct correspondence with human perceptions. We specifically focus here on haptics, but the aim is to eventually gather knowledge about all senses.
- The *Control Process* ontology aims at gathering the tasks and protocols necessary in order to perform the quality control.

The haptic quality control will make use of the elements of each of the domain ontologies, in a flexible and adapted way following the context of application and industrial constraints. The following sections focus on the *Sensory Perception* ontology, and more specifically on the part related to haptic perceptions. The other domain ontologies involved in this structure will not be extensively detailed in this paper. They are part of the development process for the formalization of haptic quality control. In particular, the *Application Context* and *Control Process* ontologies are the main parts that will enable to setup the Meta-knowledge component.

In addition, the proposed ontological structure has been made general enough in order to foresee possible utilization with other kinds of senses: for instance vision, audition or taste, as shown with dotted line boxes in Fig. 3. Indeed, the aim would be to eventually gather knowledge about all these senses, and enable perceived quality control in a more global way, taking into account the full human perception experience.

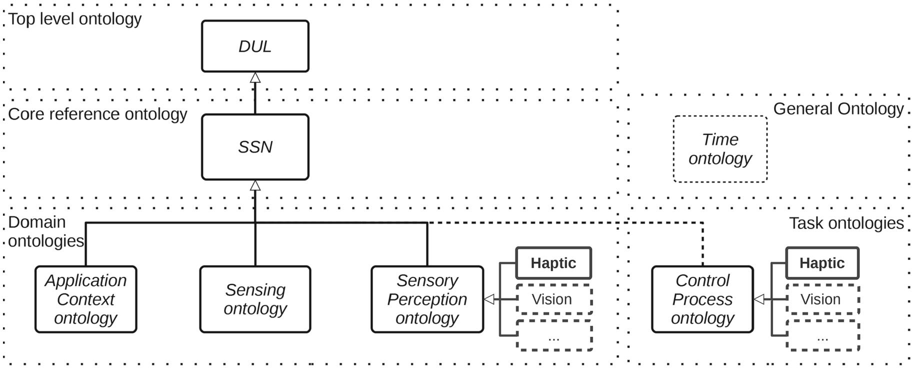


Fig. 3. General ontological structure and domain ontologies of haptic quality control.

4 Formalizing Haptic Perception Knowledge

The sense of touch has been widely studied at a biological level in order to understand how it works [14]. Regarding the description of sensations, some studies have proposed to analyse and select descriptors, but focusing on specific types of application and material, or only on specific likable features of products. A more generic and general approach is proposed in this study. It intends to formalize and structure knowledge about haptics by gathering relevant vocabulary and relations from multiple sources (such as sensory analysis studies as well as vocabulary databases). The first elements of the formalization process were presented in previous publication [12,15,16]. It mainly aimed at extracting sensation categories out of semantics in order to propose a representation as exhaustive as possible of human haptic sensations. This section presents the recent improvements of the analysis and refined models.

4.1 Perception Process

The process of generation of a haptic perception starts with touch, which can be defined as the stimulation of the skin by thermal, mechanical, chemical or electrical stimuli, combined with kinesthetic information from receptors in muscles, tendons and joints. Sensory systems, and sensory receptors in particular, activate as soon as a stimulus is detected. They transform the energy received through the stimuli into electrical energy by a change of neuronal electrical potential (transduction). Encoded information is then processed by the nervous system in order to produce sensations. Sensory systems located in the nervous system interpret these sensations by comparison to memories and known sensations. Perceptions

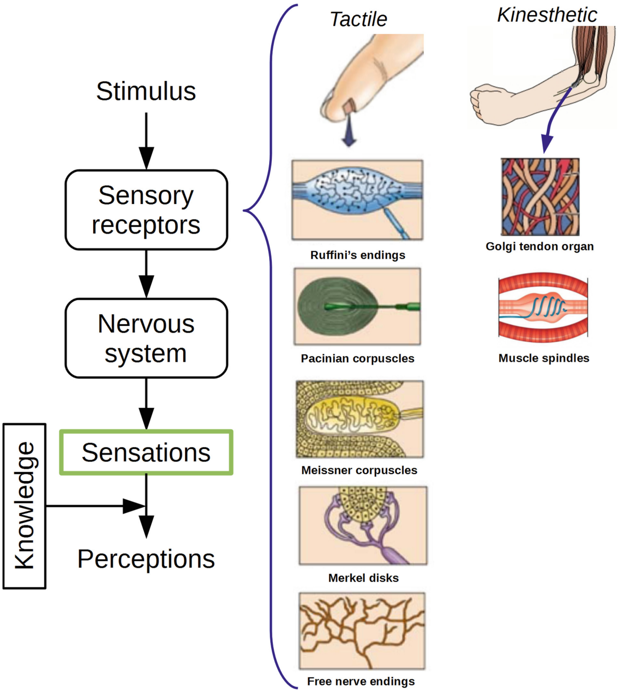


Fig. 4. The human tactile perception process. Inspired from [17,18] and [19].

are the results of this process. A schematic view of the haptic perception process is provided in Fig. 4 [17–19].

The following principle can be considered: the somatosensory system being the same, or almost the same across humans, sensations can be considered identical, or near-identical, for everybody (for similar external conditions such as temperature, humidity, level of tiredness, etc.) However, perceptions differ from one person to another, because the self-experience of a person is involved (culture, education, memories, etc.) and is very different across people. This is why we chose to focus here on sensations in order to follow an objective formalization process, while still being able to represent individual perceptions.

4.2 Usual Haptic Sensation Descriptors

Humans mainly communicate about sensations using words. Thus, the field of sensory perceptions uses a very diverse vocabulary for the description of sensations. In particular regarding haptic sensations, more than 250 descriptors could be listed when searching across the literature and dictionaries. Some examples are provided below, along with the specificities of these descriptors.

Descriptors found in the literature are usually related to specific types of products and materials. They can also depend on the language and culture of the controllers. Several methods can be listed as examples of these specificities:

– Sensotact [20] is a reference method introduced by Renault in order to describe the sensations perceived in contact of vehicle interiors. It uses ten descriptors distributed following the exploration mode (*hardness, responsiveness, memory effect, sticky, fibrous, relief, scratchy, blocking, slippery and thermal*).
– Considering textile products only, Issa et al. [21] proposed six invariant descriptors common to French and English languages (*flexible/rigid, falling, thin/thick, soft, creasable, responsive*).
– Still in the textile field of application, Picard et al. [22] found five pairs of descriptors (*soft/rough, thin/thick, mellow/hard, smooth/rough, pleasant/harsh*), from a set of twenty-four and Sola et al. [23] listed fourteen descriptors.
– Considering paper sheets, Summers et al. [24] reduced it to two descriptors (*rough* and *stiff*).
– In the field of packaging, Dumenil-Lefebvre [25] suggested two groups of descriptors, respectively for the tactile description of ground-glass (*sticky, rough, granular, slippery, cool, greasy*) and a multi-material group, including plastic, cardboard, etc. (*adherent, sticky, supple, elastic, markable, rough, granular, slippery, scratchable, cool*).

The descriptors listed in the different studies are hence very different from one product to another, as well as from one type of material to another. While some descriptors are common across different materials (for instance, the descriptor *soft* is used similarly for wood, fabric, leather, ceramic), some others are very specific to one type of material. For example, descriptors like *furry*, *fuzzy*, *fluffy*, etc. are mainly used for the description of fabric.

Translations from one language to another is also a source of variability in the way sensations are described, because the correspondence between the meanings of translated words is not always complete, or can be expressed with several different words, following the context. For example, *frais* in French could be translated into *fresh* or *cool*. Some words also do not have any translation. For instance *bouchard* in French comes from a tool: the *boucharde*, which is originally used to print marks on concrete [23]. Differences between cultures, often represented by the difference in language, can sometimes induce a difference in the meaning of the same words, for example a sensation of cold might not be perceived the same way by people living in cold or hot areas.

4.3 Formalizing Haptic Sensations

Considering the diversity of the vocabulary used in order to describe haptic sensations, formalizing the way these sensations are described has become a need when trying to build a structured representation of sensory perception knowledge. Studies performed in the context of visual quality control [26,27] demonstrated the feasibility of reducing the list of descriptors used and therefore formalizing visual aspect anomalies. A similar approach is followed here, with the addition of semantic methods, which enable to take the meaning of each word into account in order to provide a formalized representation as precise and complete as possible.

Therefore, the proposed method makes use of the semantic characteristics of the descriptors in order to extract generic categories of haptic sensations. In particular, the usual descriptors were classified. Semantic relations between descriptors were used in order to group them. Synonym and antonym links were drawn from semantic databases like Wordnet[2] and the Thesaurus[3], but also from a dictionary of sensory words [28] (in French). A graphical tool[4] and the OpenOrd method [29] were used in order to gather elements with strong relations and spread the ones with weak relations. The result is shown in Fig. 5. The OpenOrd method involves the computation of the distance between nodes (descriptors), by minimizing a formula containing attractive and repulsive terms. Thanks to this grouping we extracted categories from the meaning of the descriptors contained in these main groups.

Descriptors were also classified following three semantic axes proposed by [23]. These axes (source, effect and physical property) focus on the origins and meaning of the descriptors. They especially highlight the semantic basis of the descriptors and the links with the characteristics of a surface. The source axis refers to a perception (sensation complemented with knowledge and experience), e.g. *silky* refers to *silk*. One needs to have the knowledge of what *silk* is to understand what an *silky* sensation is. The effect axis refers to sensations that involve

[2] WordNet: https://wordnet.princeton.edu/.

[3] Thesaurus: http://www.thesaurus.com/.

[4] Gephi: Open graph visualisation platform, url: https://gephi.org/.

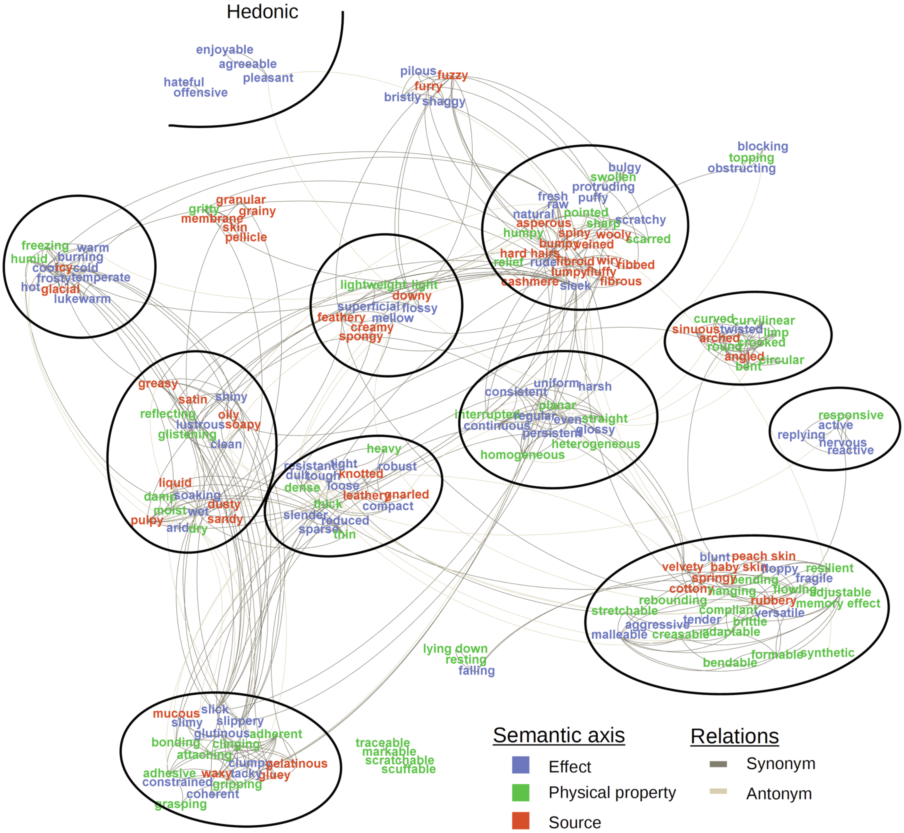


Fig. 5. Extraction of categories from usual haptic descriptors using semantic classification methods (based on [15] and [12]).

a judgement from the evaluator. These descriptors can hence be subjective, or even hedonic. Finally, the physical property axis refers to a non-hedonic sensations, which can be directly measured. This classification enabled the selection of relevant descriptors for each sensation category. Hedonic elements (which form a separate group in Fig. 5) were not selected, because they involve a strongly subjective judgment.

4.4 Elementary Haptic Sensations

A set of nine elementary haptic sensations is proposed. These elementary sensations correspond to the groups identified in the classification step - with the exception of the central group (including *homogeneous* and *heterogeneous*) corresponding to general characteristics of the other descriptors. These elementary sensations are designed to cover all haptic descriptors listed. In particular, they aim at being used in order to describe perceived sensations in a formalized manner, and can hence be used to represent the sensory descriptors.

Considering that these classes were constructed using descriptors from all kinds of domains of application, they provide a generic description of haptic sensations. This means that the descriptors found in the literature, and presented above, can be described by at least one elementary sensation. Table 1 shows the list of nine elementary sensations. Their individual definition is given. The elementary sensation are grouped into primarily tactile or primarily kinesthetic sensations, following the receptors being mainly involved in the generation of these sensations. Table 1 also shows the stimuli and exploration movements related to each elementary sensation because of their involvement in the generation of the sensation. These relations were obtained by extracting information between stimuli, exploration modes and descriptors in the following studies: [2,17,20,30–32].

Table 1. Proposed elementary sensations, and associated stimuli and exploratory movements.

Elementary sensation	Definition	Type of stimulus	Exploratory movement
Primarily tactile sensations (local impact)			
Grip	Sensation of stretching or holding back of the skin when touching the surface of an object	Stretching (and motion)	Tangential movement
Relief	Sensation of vibration or sub-centimetric shape when touching the surface of an object	Vibration or Micro shape (and motion)	Static contact or tangential movement
Hardness	Sensation of resistance of the surface of an object when pushing on it	Pressure	Orthogonal movement
Reactivity	Sensation of residual deformation or interaction with the surface of an object after pushing on it	Pressure and motion (and macro-shape)	Orthogonal and tangential movement
Residue	Sensation of matter remaining on the skin after touching the surface of an object	Persistence (and micro-shape and motion)	Tangential or orthogonal movement, or static contact
Warmth	Sensation of thermal transfer between the surface of an object and the skin when touching it	Thermal transfer	Static contact
Pain	Sensation of damaging of the skin when touching the surface of an object and its edges in particular	Damage (and motion)	Tangential or orthogonal movement, or static contact
Primarily kinesthetic sensations (global impact)			
Weight	Sensation of mass of an object in relation to its size	Mass, size	Envelopping and lifting
Shape	Sensation of global shape of the surface of an object and its edges	Macro-shape and pressure	Tangential and orthogonal movement, or envelopping

5 An Ontological Representation of the Haptic Domain

This section presents the development of a domain ontology integrating haptic perception knowledge and its alignment with Compton's core ontology, build upon the model first presented in [12] and refined here. One can make the observation that not so much existing work about the representation of sensory perception can be found in the literature and even less on haptic perception in particular. One rare example of a domain ontology representing the field of haptics was proposed by [33], but this study focused on software development for haptic interfaces. The study presented in this paper focuses instead on human perception and quality control. Thus, the proposed ontology integrates the formalization principles previously presented, and organises them so the related knowledge can be used in the proposed Smart System for Haptic Quality Control. The developed *Haptic Perception* ontology is a part of the *Sensory Perception* ontology. Both ontologies are presented in this section and the latter is detailed with an example.

5.1 General Sensory Perception Ontology

Figure 6 shows the main concepts of the proposed *Sensory Perception* ontology. This ontology is a generalization of the knowledge representations corresponding to each sense, with a general structure and the aim of being as generic as possible.

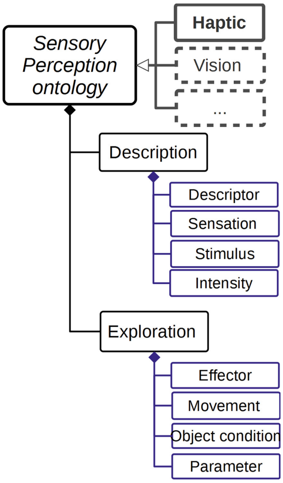


Fig. 6. Main concepts of the proposed *Sensory Perception* ontology.

The main concepts of the *Sensory Perception* ontology are the following:

- *Description* contains all the elements necessary for the description of the perception. *Description* is composed of four elements: *Descriptor, Sensation, Stimulus, Intensity.*
- *Exploration* describes the exploration modes that enable the perception. *Exploration* is composed of four elements: *Effector, Movement, Object condition, Parameter.*

- *Descriptor* integrates the usual vocabulary for describing human sensory perceptions.
- *Sensation* is a formalized description of human sensations.
- *Stimulus* is the physical phenomenon that induces sensations.
- *Intensity* describes the intensity of a sensation or a stimulus regarding a specific description.
- *Effector* is the part of the human body related to the generation of the perception.
- *Movement* is the type of movement performed in order to generate stimuli and hence the perception.
- *Object condition* describes the condition of the object when performing the exploration.
- *Parameter* lists the parameters involved in the exploration. It is directly dependent of the concepts *Effector*, *Movement* and *Object condition*.

5.2 Haptic Perception Ontology

The formalized knowledge about the haptic domain enabled to fill out the Haptic part of *Sensory Perception* ontology, with the aim of representing perceptions when touching a product.

Figure 7 shows an extract of the proposed *Sensory Perception* ontology, as well as a part of the *Sensing* ontology which is relevant here. Only a subset of the concept and relation restrictions are displayed for the sake of clarity, and in order to show a specific example of the characterisation of the descriptor *Slippery*. The

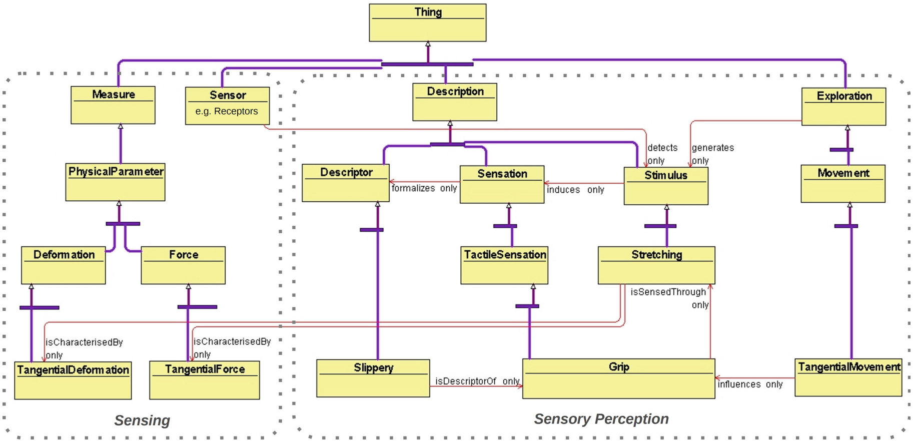


Fig. 7. Extract of the proposed *Haptic Perception* ontology. This Figure was obtained using OWLGrEd [34]. A UML-like notation is used, where boxes are OWL classes, thick lines are hierarchical relations and thin lines are restrictions, labeled with object properties. Straight segments represent the fact that certain classes have more subclasses, but only some are represented.

complete ontology includes all the elements of Table 1 as well as the full list of descriptors.

Regarding haptic perceptions in particular, the concept of *Exploration* integrates the elements necessary for the stimuli to be generated. This includes the type of effector (e.g. the hand), the movement (e.g. tangential displacement), the object condition (e.g. if it is moving or not), as well as parameters corresponding to these elements (e.g. contact duration or movement speed). The concept of *Stimulus* integrates the physical characteristics measured by human receptors, i.e. for instance *Pressure*, *Stretching*, *Vibration*, etc. The concept of *Sensation* integrates the proposed formalization of haptic sensations previously presented. The concept *Sensation* is distributed on the two classes *TactileSensation* and *KinestheticSensation.* The usual descriptors are grouped under the concept *Descriptor* and are related to the corresponding sensation categories.

The concepts of *Measure* and *Sensor* are part of the *Sensing* ontology. Figure 7 integrates them in the purpose of providing an example of the extensions of the relations previously presented.

5.3 Example: Representation of the Descriptor *Slippery*

Figure 7 serves also as an example of the relations involved in the description and characterization of the descriptor *Slippery.* First, the process of formalization of haptic knowledge previously presented provides information on relations between usual descriptors and the proposed elementary sensations. Considering this specific example case, it is possible to establish that the elementary *Sensation* of *Grip* uniquely describes the *Descriptor Slippery. Grip* is perceived through the *Stimulus Stretching.* As seen above, the elementary sensation of Grip is also related to a tangential movement, which is represented in the ontology by a restriction between *Grip* and the *Movement TangentialMovement.* Moreover, this specific type of *Stimulus* is characterized by the physical parameters *TangentialDeformation* and *TangentialForce.*

This representation provides hence all the elements necessary to link haptic sensations - and even more specifically usual descriptors of these sensations - to the physical properties of the stimuli that can be measured in order to provide intensity values of haptic sensations. Moreover, this ontology also provides the related exploration elements, as well as the sensory receptors (through the *Sensing* ontology) involved in the perception process. Note again that many more relations exist between these concepts, for instance between *Movement* and *Stimulus*, but for the sake of clarity, only the most relevant to this example were shown in Fig. 7.

5.4 Alignment to the SSN Core Ontology

With regards to the aim of applying haptic perception knowledge to quality control, the higher level *SSN* ontology brings a formal structure to the proposed model and will enable to reason about it in a more general and coherent manner. Indeed, *SSN* is designed to provide a formal ontological framework in order to represent the interactions between sensors and properties, or more generally between a sensing system and features of interest. The alignment to a part of the *SSN* ontology, and more specifically to the sensor-stimulus-observation pattern (shown in Fig. 2), is relatively natural considering the concepts proposed in the *Sensory Perception* ontology. The following alignment is proposed, with SP = *Sensory Perception* and S = *Sensing*:

$$\begin{aligned} SP : Sensation &\sqsubseteq SSN : Property \\ SP : Descriptor &\sqsubseteq SSN : FeatureOfInterest \\ SP : Stimulus &\sqsubseteq SSN : Stimulus \\ SP : Exploration &\sqsubseteq SSN : Sensing \\ S : PhysicalParameter &\sqsubseteq SSN : Property \\ S : Sensor &\sqsubseteq SSN : Sensor \end{aligned}$$

In particular, the SSN concept *SSN:Property* is defined as an observable characteristic of real-world entities (*SSN:FeatureofInterest*), which are not directly observable. *SP:Descriptor* can be aligned to *SSN:FeatureofInterest*, because it corresponds to the way people usually communicate about sensations, which is not directly observable. At the opposite, *SP:Sensation* and *S:PhysicalParameter* correspond to observable characteristics of *SP:Descriptor*. They can hence be aligned to *SSN:Property*.

The concept *Sensor* can be aligned to *SSN:Sensor* because it detects stimuli. Moreover, in SSN, *SSN:Sensor* implements *SSN:Sensing*, and considering human sensory perception, *SP:Exploration* can be considered as a way to implement sensory receptors (*S:Sensor*, which hence corresponds to *SSN:Sensing*. The proposed ontology adds the fact that some *SP:Exploration* concepts also enable the generation of *SP:Stimulus* through contact and movement.

In addition, some concepts from the *Application Context* ontology can be matched with the concept of *SSN:Observation* which represents the conditions of observation. However the SSN ontology does not include sufficient details about context. This is why future work within this project will be focusing on interaction with context, and in particular in order to provide a complete description of the *Meta-knowledge* component of the KREM framework.

6 Conclusions and Perspectives

This work provides the first steps of the development of a Smart System for haptic quality control. A general framework was presented, based on the KREM model which provides a separation of the modules involved in the development of this system, while still enabling interaction between them. The main aim in using this framework is to allow for more flexibility in the development, considering future applications of the system, in particular regarding the integration of experience and enrichment of the knowledge base.

A global ontological structure was defined in order to fulfil the needs for knowledge integration from the multiple domains involved in this project, i.e.

human perception, quality control, instrumentation and industrial context of application. Moreover, the structuring of the knowledge was founded on well-known high level ontologies, and the proposed domain ontology for the representation of sensory perception knowledge was aligned with the SSN ontology in particular.

The construction of this ontology involved the formalization of the knowledge related to haptic perceptions. A generic description was presented with the proposition of elementary haptic sensations, based on a semantic analysis of a large set of sensory vocabulary. Through relations found across the literature between sensory descriptors, stimuli and exploration modes, a model of correspondence was proposed and implemented in the ontology. Reasoning on this model will provide the Smart System with expert knowledge about sensory perception, and enable manufacturers to standardize haptic quality control while skipping relatively long and heavy usual sensory analysis processes.

This paper focused on the *Knowledge* component of the KREM framework. The next steps in the development of the propose Smart System will include the *Meta-knowledge* and *Rules* modules which will enable to reason according to the context of application and hence adapt the selected quality control methods to manufacturing contexts (e.g. adapting the selection of haptic sensations and exploration parameters to the material of a product). In addition, current industrial testings of the proposed models will provide evaluation results and enable continuous enrichment of the knowledge base. Furthermore, the proposed system is intended to eventually provide an automated process of perceived quality control. Knowledge about sensors and products will hence also be explored, as well as the relations between data from the sensors and haptic sensations. This corresponds to the problem of symbol anchoring and brings further perspectives to this study.

Acknowledgments. This work has been done within a thesis project funded by the French technological research association (ANRT) as well as the company INEVA (INEVA: http://www.ineva.fr/). This work is the result of a collaboration between three parties, which are all acknowledged here: the company INEVA, the INSA de Strasbourg (with the ICube laboratory) and the University of Savoie Mont Blanc (with the SYMME laboratory).

References

1. Katz, D.: The World of Touch. Psychology Press, New York (2013)
2. Lederman, S., Klatzky, R.L.: Haptic perception: a tutorial. Atten. Percept. Psychophys. **71**, 1439–1459 (2009)
3. Zanni-Merk, C.: KREM: a generic knowledge-based framework for problem solving in engineering - proposal and case studies. In: INSTICC (eds.) 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, pp. 381–388. Science and Technology Publications, Lda (2015)

4. Zanni-Merk, C., Marc-Zwecker, S., Wemmert, C., de Beuvron, B.F.: A layered architecture for a fuzzy semantic approach for satellite image analysis. Int. J. Knowl. Syst. Sci. **6**, 31–56 (2015)
5. Gartiser, N., Zanni-Merk, C., Boullosa, L., Casali, A.: A semantic layered architecture for analysis and diagnosis of SME. Procedia Comput. Sci. **35**, 1165–1174 (2014). https://doi.org/10.1016/j.procs.2014.08.212. Elsevier Masson SAS
6. Milton, N.: Knowledge Technologies. Polimetrica, Milano (2008)
7. Sanín, C., Szczerbicki, E.: Experience-based knowledge representation: SOEKS. Cybern. Syst. **40**, 99–122 (2009)
8. Aamodt, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Commun. **7**, 39–59 (1994)
9. Dey, A.K., Abowd, G.D., Salber, D.: A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Hum.-Comput. Interact. **16**, 97–166 (2001)
10. Compton, M., et al.: The SSN ontology of the W3C semantic sensor network incubator group. J. Semant. **17**, 25–32 (2012)
11. DOLCE+DnS_Ultralite (2010). ontologydesignpatterns.org
12. Albert, B., et al.: A smart system for haptic quality control - introducing an ontological representation of sensory perception knowledge. In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Porto. SCITEPRESS - Science and Technology Publications, pp. 21–30 (2016)
13. Roussey, C., Pinet, F., Kang, M.A., Corcho, O.: An introduction to ontologies and ontology engineering. In: Falquet, G., Métral, C., Teller, J., Tweed, C. (eds.) Ontologies in Urban Development Projects. AI&KP, vol. 1, pp. 9–38. Springer, London (2011). https://doi.org/10.1007/978-0-85729-724-2_2
14. Hayward, V.: Is there a ‘plenhaptic’ function? Philos. Trans. Royal Soc. B: Biol. Sci. **366**, 3115–3122 (2011)
15. Albert, B., et al.: Generic and structured description of tactile sensory perceptions. In: KEER, Number September, Leeds, UK (2016)
16. Albert, B., et al.: Formalisation du Contrôle Qualité Haptique: Structuration Sémantique des Sensations Haptiques. IC (2016)
17. De Boissieu, F.: Toucher artificiel à base d’un microcapteur d’effort: traitement du signal et des informations associées. Ph.D. thesis, Université de Grenoble (2010)
18. De Rossi, D., Scilingo, E.P.: Skin-like sensor arrays. Encycl. Sens. **10**, 1–22 (2006)
19. Purves, D., et al.: Neuroscience, 2nd edn. Sinauer Associates, Sunderland (2001)
20. Crochemore, S., Vergneault, C., Nesa, D.: A new reference frame for tactile perceptions: sensotact. 5th Rose Mary Pangborn, Boston, MA, USA, pp. 20–24 (2003)
21. Issa, M., Schacher, L., Adolphe, D.C.: Invariant attributes in the tactile characterization of fabrics. In: Proceeding of the Fiber Society Spring Conference (2005)
22. Picard, D., Dacremont, C., Valentin, D., Giboreau, A.: Perceptual dimensions of tactile textures. Acta Psychol. **114**, 165–184 (2003)
23. Sola, C.: Y a pas de mots pour le dire, il faut sentir: Décrire et dénommer les happerceptions professionnelles. Terrain (2007)
24. Summers, I.R., Irwin, R.J., Brady, A.C.: Haptic discrimination of paper. In: Human Haptic Perception: Basics and Applications, Number December 2007, pp. 525–535. Birkhäuser Basel (2007)

25. Dumenil-Lefebvre, A.: Integration Des Aspects Sensoriels Dans La Conception Des Emballages En Verre: Mise Au Point D'Un Instrument Methodologique À Partir Des Techniques D'Evaluation Sensorielle. Ph.D. thesis, Ecole nationale supérieure d'arts et métiers (2006)
26. Baudet, N.: Maîtrise de la qualité visuelle des produits - Formalisation du processus d'expertise et proposition d'une approche robuste de contrôle visuel humain. Ph.D. thesis, Université de Grenoble (2012)
27. Maire, J.L., Pillet, M., Baudet, N.: Measurement of the perceived quality of a product - characterization of aesthetic anomalies. Int. J. Metrol. Qual. Eng. **4**, 63–69 (2013)
28. Bassereau, J.F., Charvet-Pello, R.: Dictionnaire des mots du sensoriel. Lavoisier (2011)
29. Martin, S., Brown, W.M., Klavans, R., Boyack, K.W.: OpenOrd: an open-source toolbox for large graph layout. In: Proceedings of the SPIE, vol. 7868, pp. 786806–786811 (2011)
30. Bensmaïa, S., Hollins, M.: Pacinian representations of fine surface texture. Percept. Psychophys. **67**, 842–854 (2005)
31. Jones, L.A., Lederman, S.J.: Human Hand Function. Oxford University Press, Oxford (2006)
32. Hatwell, Y., Streri, A., Gentaz, E.: Touch for Knowing. John Benjamins, Amsterdam (2003)
33. Myrgioti, E., Bassiliades, N., Miliou, A.: Bridging the HASM: an OWL ontology for modeling the information pathways in haptic interfaces software. Expert Syst. Appl. **40**, 1358–1371 (2013)
34. Bārzdiņš, J., Bārzdiņš, G., Čerāns, K., Liepiņš, R., Sproģis, A.: UML style graphical notation and editor for OWL 2. In: Forbrig, P., Günther, H. (eds.) BIR 2010. LNBIP, vol. 64, pp. 102–114. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16101-8_9

The Mereologies of Upper Ontologies

Lydia Silva Muñoz[1] and Michael Grüninger[2](✉)

[1] Computer Science, University of Toronto,
40 St. George Street, Toronto, ON M5S 2E4, Canada
silva@cs.toronto.edu

[2] Mechanical and Industrial Engineering, University of Toronto,
5 King's College Road, Toronto, ON M5S 3G8, Canada
gruninger@mie.utoronto.ca

Abstract. Mereology, the formal theory of parts and wholes, has a played a prominent role within applied ontology. As a fundamental set of concepts for commonsense reasoning, it also appears in a number of upper level ontologies. Furthermore, such upper-level ontologies provide an account of the most basic, domain-independent, existing entities, such as time, space, objects, and processes. In this paper, we verify the core characterization of mereologies of the Suggested Upper Merged Ontology (SUMO), and the mereology of the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), while relating their axiomatizations via ontology mapping. We show that the existing axiomatization of SUMO omits some of the intended models of classical mereology, and we propose the correction and addition of axioms to address this issue. In addition, we show the formal relationship between the axiomatization of mereology in both upper-level ontologies.

Keywords: DOLCE · DOLCE-CORE · SUMO · Ontology mapping
Ontology verification · Upper-level ontology · Mereology · Topology
Mereotopology

1 Introduction

Automatic applications appealing to ontologies for interoperation are unambiguously integrated only when the models of their shared features are equivalent. However, ontologies admitting unintended models ambiguously characterize their vocabularies, which can generate misunderstandings that hinder interoperability.

Upper-level ontologies, also called *foundational* ontologies, provide an account of the most basic, domain independent, existing entities, such as time, space, objects and processes. As ontologies are crucial for the Semantic Web, upper level ontologies are essential for the ontology engineering cycle in activities such as ontology building and integration. Upper level ontologies can be used as the foundational substratum on which new ontologies are developed, because they provide some fundamental ontological distinctions, which can help

The original version of this chapter was revised: The family name of the first author was incorrect. The correction to this chapter is available at https://doi.org/10.1007/978-3-319-99701-8_19

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 191–212, 2019.
https://doi.org/10.1007/978-3-319-99701-8_9

the designer in her task of conceptual analysis, [11]. They can be used as a backbone on top of which more specific concepts can be characterized while reusing their root vocabulary and their general knowledge. In ontology integration, they can be used as oracles for meaning clarification [6].

Upper-level ontologies are expected to be mostly consulted for meaning negotiation and not for terminological reasoning, therefore they have to be represented in expressive languages that can convey every feature of the characterized entities. However, less accurate representations in lightweight languages should be available too, which can be extended with domain specific axiomatizations when the use of lightweight languages is necessary.

Since foundational ontologies are expected to be broadly reused and extended, they are not expected to admit unintended models which the expressivity of their representation language can rule out, and it is also expected that they do not miss intended models. If a upper-level ontology does not admit those models which it is expected to admit, misunderstandings among applications who subscribe to their account of the world can occur.

Various upper level ontologies have been developed in languages with higher or equivalent expressivity to first-order logic, such as SUMO [18] and DOLCE [2,8], and translations of them, with loss, to lightweight language OWL[1], made available. Therefore, semantic mappings connecting their axiomatizations are necessary to facilitate interoperability among applications that commit to the characterizations provided by different upper level ontologies. Those mappings need to be formal, which guarantees their interpretability by automatic agents, and also need to be represented in an expressive language such as standard first-order logic.[2]

Ontology verification [10] is the process by which a theory is checked to rule out unintended models, and possibly characterize missing intended ones. Therefore, ontology verification reduces semantic ambiguity. Since foundational ontologies are expected to be broadly reused, their verification results necessary.

In this paper[3], we verify the subtheory of core mereotopological concepts of the SUMO foundational ontology and the mereology of the DOLCE-CORE, the fragment of DOLCE focused on entities that exist on time. In addition, we formally relate their respective axiomatizations via first-order logic mappings. As a result, we propose the correction, and addition, of some axioms which rule out unintended models or characterize missing ones. As an additional outcome of our work, we have produced a modular representation stated in standard

[1] https://www.w3.org/2001/sw/wiki/OWL.

[2] The expressive power of first-order logic makes its use necessary for the representation of mappings that characterize features that are not representable in lightweight languages, such as Description Logics. In addition, checking the correctness of those mappings results facilitated by the fact that first-order theorem proving in standard first-order logic is a mature field, and, although semi-decidable, first-order reasoning on small modules results in an acceptable trade-off among expressivity and efficiency.

[3] This paper is an extended and expanded version of the paper "Verifying and Mapping the Mereotopology of Upper-Level Ontologies" that originally appeared in the Proceedings of Knowledge Engineering and Ontology Design (KEOD) 2016 [16].

first-order logic of the complete SUMO subtheory of mereotopology. We have used automatic theorem prover Prover9 and model finder Mace4 [15] for the automatic tasks involved in the work described in this paper.

2 Ontology Mapping and Verification

Ontology mapping, also called *ontology matching*, and *ontology alignment*, is concerned with the explicit representation of the existing semantic correspondences among the axiomatizations of different ontologies[4] via *bridge axioms* [6], which are called *translations definitions* in the context of first-order logic.

Building a map between two first-order logic ontologies T_1 and T_2 that interprets the first into the second involves translating every symbol of theory T_1 into the language of T_2, translating every sentence of T_1 into the language of T_2, and checking the ability of T_2 to entail every axiom of T_1. The following definition formalizes the notion of relative interpretation between first-order logic theories.

Definition 1. *A map π interprets a theory T_1 into a theory T_2 iff for every sentence α in the language of T_1, $T_1 \models \alpha \Rightarrow T_2 \models \alpha^{\pi}$; being α^{π} the syntactic translation of α into the language of T_2.*

The following theorem that follows fom [5], introduces a fundamental relation between the models of a theory and the models of the theories that it interprets. Given such a relation, in order to demonstrate that a given theory T_2 can represent every feature that another theory T_1 represents, it suffices to demonstrate that theory T_2 is able to interpret theory T_1.

Theorem 1. *If a theory T_1 is interpreted by a theory T_2 by means of a given map π, there is another map δ that sends every model of T_2 into a model of T_1.*

An ontology admits unintended models when it is possible to find features of its underlying conceptualization which are not characterized by its axiomatization. *Ontology Verification* in first-order logic [10] is based on the fact that theories with different vocabularies unambiguously characterize the same concepts only if their sets of models are equivalent. Verifying an ontology T ideally consists of classifying the actual models $\mathfrak{M}$ of T by means of a *representation theorem*,[5] which relates the models of T with the models $\mathfrak{M}^{intended}$ of an alternative axiomatization of T built with well understood theories. Such a representation theorem must be either proved or disproved. The following definition from [19] relates the notion of *ontology mapping* with the fundamentals of *ontology verification*:

[4] We assume that an ontology is a set of sentences called *axioms* closed under logical entailment that state the properties that characterize the behaviour of a set of symbols representing constants, relations and functions, called the *signature* of the ontology.

[5] A *representation theorem* is a theorem that formally classifies a given class of structures as equivalent to another class of structures whose properties are better understood. The stated equivalence makes possible the extrapolation of those properties to the classified structures, facilitating their understanding.

Definition 2. *Two theories T_1 and T_2 are* ***synonymous*** *iff there exist two sets of translation definitions Δ and Π, respectively from T_1 to T_2 and from T_2 to T_1, such that $T_1 \cup \Pi$ is logically equivalent to $T_2 \cup \Delta$.*

Given Definition 2, from Theorem 1 follows that the models of synonymous theories are equivalent, and therefore *ontology mapping* can be used for classifying the sets of models of two ontologies as equivalent.

3 SUMO

SUMO [18] is a freely available upper level ontology intended to describe the world as perceived by humans, based on human knowledge and culture, in opposition to *ontological realism* [9], which is meant to present the world as it is, independently of the bias of human perception. In addition to the main ontology, which contains about 4000 axioms, SUMO has been extended with a mid-level ontology and a number of domain specific ontologies, all of which account for 20,000 terms and 70,000 axioms. SUMO has been translated into OWL and WordNet [17]. The representation language of SUMO is SUO-KIF[6], a very expressive dialect of KIF[7] with many-sorted features, whose syntax permits higher-order constructions such as predicates that have other predicates, or formulas, as their arguments, and the existence of predicates and functions of variable arity [1].

We have translated (with loss) into standard first-order logic, and modularized, the subset of SUMO that characterizes the notion of mereotopology, which resulted in the hierarchy of subtheories shown in Fig. 1, where each theory conservatively extends[8] its related theories below. Due to space limitations, we only address in this work the study of modules T_{part}, T_{sum}, $T_{product}$, $T_{decomposition}$, $T_{topology}$, and $T_{mereotopology}$. The first-order logic axiomatization of all the modules shown in Fig. 1 can be found at colore.oor.net/ontologies/sumo/modules.

SUMO adopts various partial orderings to address the part-whole relationship in different categories. Regarding entities that are in space and time, classified as *Physical* in SUMO, relations *part* and *subProcess* respectively characterize part-whole relations for members of *Object* and *Process*, while relation *temporalPart* represents part-whole for members of *TimePosition*, which extends to points and intervals of time.

3.1 The Subtheory T_{part}

The subtheory T_{part} represents the relation among a whole and its parts by axiomatizing the primitive relation *part* and using conservative definitions for the *overlapsSpatially*, *overlapsPartially*, and *properPart* relations. Extracting

[6] http://suo.ieee.org/SUO/KIF/suo-kif.html.

[7] http://logic.stanford.edu/kif/kif.html.

[8] A theory T' is a *conservative extension* of a theory T if every theorem of T is a theorem of T', and every theorem of T' in the signature of T is also a theorem of T.

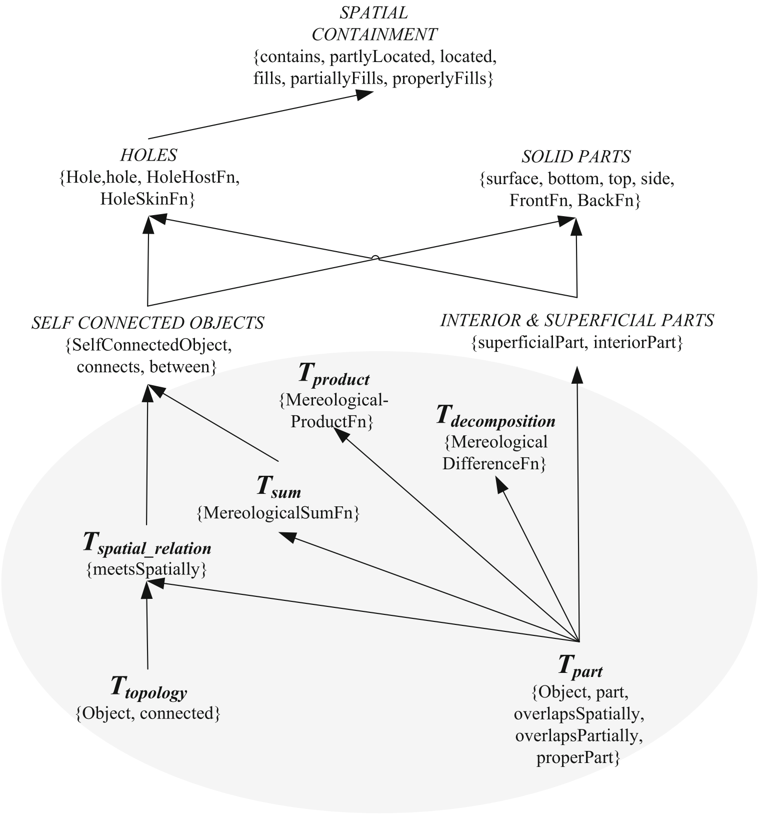

SPATIAL CONTAINMENT
{contains, partlyLocated, located, fills, partiallyFills, properlyFills}
HOLES
{Hole,hole, HoleHostFn, HoleSkinFn}
SOLID PARTS
{surface, bottom, top, side, FrontFn, BackFn}
SELF CONNECTED OBJECTS
{SelfConnectedObject, connects, between}
INTERIOR & SUPERFICIAL PARTS
{superficialPart, interiorPart}
T product
{Mereological-ProductFn}
T decomposition
{Mereological DifferenceFn}
T sum
{MereologicalSumFn}
T spatial_relation
{meetsSpatially}
T topology
{Object, connected}
T part
{Object, part, overlapsSpatially, overlapsPartially, properPart}

Fig. 1. Modular decomposition of the SUMO axiomatization of concepts related to mereotopology. Theories in the shaded region are discussed in this paper; the remaining subtheories of SUMO are left for future work. Arrows point to conservative extensions among modules. Signature members are shown in the module that first introduces them. (Original figure from [16].)

the sentences from SUMO that use the primitive signature $\{Object, part\}$, we obtain the following subtheory:

Definition 3. *T_{part} is the subtheory composed by axioms (1) to (7).*

$$(\forall x, y) part(x, y) \rightarrow Object(x) \wedge Object(y) \tag{1}$$

$$(\forall x) Object(x) \rightarrow part(x, x) \tag{2}$$

$$(\forall x, y) part(x, y) \wedge part(y, x) \rightarrow (x = y) \tag{3}$$

$$(\forall x, y, z) part(x, y) \wedge part(y, z) \rightarrow part(x, z) \tag{4}$$

$$(\forall x, y) overlapsSpatially(x, y) \leftrightarrow (\exists z (part(z, x) \wedge part(z, y))) \tag{5}$$

$$(\forall x, y) overlapsPartially(x, y) \leftrightarrow \neg part(x, y) \\ \wedge \neg part(y, x) \wedge (\exists z) part(z, x) \wedge part(z, y) \tag{6}$$

$$(\forall x, y) properPart(x, y) \leftrightarrow part(x, y) \wedge \neg part(y, x) \tag{7}$$

The first question we need to address is whether or not this subtheory is a module of SUMO, or whether there are additional sentences in the signature $\{Object, part\}$ that are entailed by the remaining axioms of SUMO. Before we can fully answer this question, we need to consider the other subtheories of SUMO that are related to mereology.

3.2 Subtheory T_{sum}

The mereological sum of two parts into a whole is represented in the subtheory T_{sum} by the function symbol $MereologicalSumFn$.

Definition 4. *The subtheory T_{sum} is the subtheory that extends T_{part} in the expanded signature $\{Object, part, MereologicalSumFn\}$ by means of axioms (8) and (9).*

$$(\forall x, y, z) Object(x) \wedge Object(y) \rightarrow \\ ((z = MereologicalSumFn(x, y)) \rightarrow (\forall p)(part(p, z) \leftrightarrow (part(p, x) \vee part(p, y))) \tag{8}$$

$$(\forall x, y) Object(x) \wedge Object(y) \rightarrow Object(MereologicalSumFn(x, y)) \tag{9}$$

Given two objects, the existence of their mereological sum is guaranteed in this theory due to the use of a function to represent such an operation (since functions in first-order logic are total).

We can immediately find some straightforward consequences[9] of the axioms of T_{sum}:

Proposition 1.

$$T_{sum} \models (\forall x, y, z) Object(x) \wedge Object(y) \wedge \\ (z = MereologicalSumFn(x, y)) \rightarrow part(z, x) \vee part(z, y) \tag{10}$$

$$T_{sum} \models (\forall x, y, z) Object(x) \wedge Object(y) \wedge \\ (z = MereologicalSumFn(x, y)) \rightarrow part(x, z) \wedge part(y, z) \tag{11}$$

Given theorems (10) and (11), and due to the antisymmetry of relation *part*, it holds that z must be x or y, this fact entails that every pair of objects in the universe of every interpretation of SUMO must be in relation *part*, which shows that SUMO omits models where there exist objects that are disjoint, or that overlap without being one part of the other, as depicted in parts *(b)* and *(c)* of Fig. 2. The following proposition formalizes this claim:

[9] The proofs for all Propositions have been found using the Prover9 automated theorem prover, and models were constructed using Mace4. Results are available at: colore.oor.net/ontologies/sumo/mereotopology/proofs.

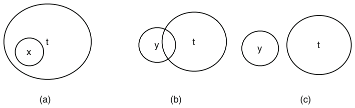


Fig. 2. With the original characterization of mereological sum, every two objects in every model of SUMO must be in relation *part*, such as objects x and y in *(a)*. Models corresponding to *(b)* and *(c)* with overlapping objects without being one part of the other, or with disjoint objects, are not admitted by SUMO submodule T_{sum}. (Original figure from [16].)

Proposition 2. $T_{sum} \models (\forall x,y) Object(x) \wedge Object(y) \rightarrow part(x,y) \vee part(y,x)$.

In other words, SUMO entails that the *part* relation is synonymous with a linear ordering. Although there is much discussion in the philosophical and applied ontology literature [20] over which axioms should constitute a mereology, there is nobody who proposes that all models of an axiomatization of mereology be synonymous with linear orderings. In order to allow those omitted models that Proposition 2 identifies, we propose a modification of T_{sum}:

Definition 5. *$T_{extended\_sum}$ is the theory which extends T_{part} with the sentences*

$$(\forall x,y,z) Object(x) \wedge Object(y) \rightarrow \\ ((z = MereologicalSumFn(x,y) \rightarrow (\forall p)(part(z,p) \leftrightarrow part(x,p) \wedge part(y,p))) \tag{12}$$

$$(\forall x,y) Object(x) \wedge Object(y) \rightarrow Object(MereologicalSumFn(x,y)) \tag{13}$$

The following proposition shows that $T_{extended\_sum}$ does not rule out models in which overlapping or disjoint objects exist.

Proposition 3.

$$T_{extended\_sum} \not\models (\forall x,y) Object(x) \wedge Object(y) \rightarrow (part(x,y) \vee part(y,x)$$

If we look more closely at the proposed axiomatization, we can see that the $MereologicalSumFn$ function is commutative and idempotent:

Proposition 4.

$$T_{extended\_sum} \models (\forall x,y,z) Object(x) \wedge Object(y) \wedge Object(z) \wedge$$

$$(MereologicalSumFn(x,y) = z) \rightarrow (MereologicalSumFn(y,x) = z)$$

$$T_{extended\_sum} \models (\forall x,y,z) part(x,y) \rightarrow (MereologicalSumFn(x,y) = y)$$

This leads us to consider how $T_{extended\_sum}$ is related to lattice theory [4].

Theorem 2. *$T_{extended\_sum}$ is synonymous with $T_{join\_semilattice}$*[10].

[10] colore.oor.net/ontologies/lattices/join_semilattice.clif.

Proof. Let Δ be the sentence

$$(\forall x,y,z)\ (MereologicalSumFn(y,x) = z) \leftrightarrow (join(x,y) = z)$$

Using Prover9, we can show that $T_{extended\_sum} \cup \Delta \models T_{join\_semilattice}$, and $T_{join\_semilattice} \cup \Delta \models T_{extended\_sum}$ □

We can also specify an extension of T_{part} in the same signature.

Definition 6. *T_{part_sum} is the extension of T_{part} with the sentence*

$$(\forall x,y)(\exists j)\ (part(x,j) \wedge part(y,j) \wedge ((\forall z)\ (part(x,z) \wedge part(y,z) \supset part(j,z))))$$

Theorem 3. *T_{part_sum} is synonymous with $T_{strong\_lub\_mereology}$*[11]

Proof. To disambiguate the signatures of SUMO and other existing mereologies, let $part^{s}umo(x,y)$ be the relation in the signature of SUMO.

Let Δ be the sentence

$$(\forall x,y)\ part^{sumo}(x,y) \leftrightarrow part(x,y))$$

Using Prover9, we can show that $T_{part\_sum} \models T_{strong\_lub\_mereology}$, and $T_{strong\_lub\_mereology} \models T_{part\_sum}$ □

It should be noted that the axiomatization of $T_{strong\_lub\_mereology}$ corresponds to the sentence SA13 in [20].

Proposition 5.

$$SUMO \models T_{part\_sum}$$

3.3 Subtheory $T_{product}$

Given two objects, their mereological product intuitively corresponds to their intersection. SUMO represents the notion of mereological product by means of the function $MereologicalProductFn$.

Definition 7. *$T_{product}$ is the subtheory of SUMO that extends theory T_{part} in the expanded signature $\{Object, part, MereologicalProductFn\}$ by the sentences:*

$$\begin{gathered}(\forall x,y,z) Object(x) \wedge Object(y) \rightarrow \\ ((z = MereologicalProductFn(x,y)) \rightarrow (\forall p)(part(p,z) \leftrightarrow part(p,x) \wedge part(p,y)))\end{gathered} \tag{14}$$

$$(\forall x,y) Object(x) \wedge Object(y) \rightarrow Object(MereologicalProductFn(x,y)) \tag{15}$$

[11] colore.oor.net/ontologies/mereology/strong_lub_mereology.clif.

Given two objects, the existence of their mereological product is guaranteed due to the use of a function to represent such an operation.

The characterization of mereological product in SUMO corresponds to the *infimum* or *meet* of the corresponding arguments on the lattice that relation *part* defines. We have found that from the characterization of mereological product of SUMO follows that every pair of objects must overlap, which indicates that SUMO omits those models where there exist objects that do not overlap (in other words, all elements overlap each other in all models of SUMO):

Proposition 6. $T_{product} \models (\forall x, y) Object(x) \wedge Object(y) \rightarrow (overlapsSpatially(x, y))$.

In order to allow models in which nonoverlapping elements exist, we propose a modification of SUMO:

Definition 8. *$T_{extended\_product}$ is the theory which extends T_{part} by the sentences*

$$(\forall x, y, z) overlapsSpatially(x, y) \rightarrow \\ ((z = MereologicalProductFn(x, y)) \rightarrow (\forall p)(part(p, z) \leftrightarrow part(p, x) \wedge part(p, y))) \quad (16)$$

$$(\forall x, y) Object(x) \wedge Object(y) \rightarrow Object(MereologicalProductFn(x, y)) \quad (17)$$

The following propositions show that $T_{extended\_product}$ does not rule out models in which there exist nonoverlapping objects, and that $MereologicalProductFn$ is commutative and idempotent.

Proposition 7.

$$T_{extended\_product} \not\models (\forall x, y) Object(x) \wedge Object(y) \rightarrow overlapsSpatially(x, y)$$

Proposition 8.

$$T_{extended\_product} \models (\forall x, y, z) Object(x) \wedge Object(y) \wedge Object(z) \wedge \\ (MereologicalProdFn(x, y) = z) \rightarrow (MereologicalProdFn(y, x) = z)$$

$$T_{extended\_product} \models (\forall x, y, z) part(x, y) \rightarrow (MereologicalProdFn(x, y) = x)$$

We can also specify an extension of T_{part} in the same signature.

Definition 9. *T_{part_prod} is the extension of T_{part} with the sentence*

$$(\forall x, y)\ overlapsSpatially(x, y) \supset (\exists z)\ ((\forall u)\ (part(u, z) \leftrightarrow (part(u, x) \wedge part(u, y))))$$

Calling $part^{sumo}$ to the relation *part* of T_{part}, the following property holds:

Theorem 4. *T_{part_prod} is synonymous with $T_{prod\_mereology}$*[12]*.*

Proof. Let Δ be the sentence

$$(\forall x, y)\ part^{sumo}(x, y) \leftrightarrow part(x, y))$$

Using Prover9, we can show that $T_{part\_prod} \models T_{prod\_mereology}$, and $T_{prod\_mereology} \models T_{part\_prod}$ □

[12] colore.oor.net/ontologies/mereology/prod_mereology.clif.

3.4 Subtheory $T_{decomposition}$

The remainder between a whole and its proper parts is represented by the function *MereologicalDifferenceFn*.

Definition 10. *$T_{decomposition}$ is the subtheory of SUMO that extends T_{part} by the sentences*

$$(\forall x, y, z) Object(x) \land Object(y) \to ((z = MereologicalDifferenceFn(x, y)) \to \\ (\forall p) properPart(p, z) \leftrightarrow properPart(p, x) \land \neg properPart(p, y)) \quad (18)$$

$$(\forall x, y) Object(x) \land Object(y) \to Object(MereologicalDifferenceFn(x, y)) \quad (19)$$

Because the mereological difference, or remainder, between a whole and one of its parts is represented in SUMO by a function, its existence is guaranteed in every case at the expenses of having arbitrary values of the function $MereologicalDifferenceFn$.

We have found that the axiomatization of *MereologicalDifferenceFn* given by (18) and (19) entails an unusual result in which the remainder overlaps with the subtrahend:

Proposition 9.

$$T_{decomposition} \models (\forall x, y, z) Object(x) \land Object(y) \\ \land (z = MereologicalDifferenceFn(x, y)) \land properPart(y, x) \to properPart(y, z))$$

In order to eliminate such a class of unintended models, we propose the following modification, and prove by means of Proposition 10 that the new theory does not admit these models.

Definition 11. *$T_{extended\_decomp}$ is the theory that extends T_{part} with the following sentences*

$$(\forall x, y, z) Object(x) \land Object(y) \to ((MereologicalDifferenceFn(x, y) = z) \to \\ (\forall p)(part(p, z) \leftrightarrow part(p, x) \land \neg overlapsSpatially(p, y))) \quad (20)$$

$$(\forall x, y) Object(x) \land Object(y) \to Object(MereologicalDifferenceFn(x, y)) \quad (21)$$

Proposition 10.

$$T_{extended\_decomp} \not\models (\forall x, y, z) Object(x) \land Object(y) \\ \land (z = MereologicalDifferenceFn(x, y)) \land properPart(y, x) \to properPart(y, z))$$

We can also specify an extension of T_{part} in the same signature.

Definition 12. *T_{part_decomp} is the extension of T_{part} with the sentence*

$$(\forall x, w)\ \neg part(w, x) \supset (\exists z)\ ((\forall y)\ (part(y, z) \leftrightarrow \neg overlapsSpatially(y, x)))$$

Theorem 5. *T_{part_decomp} is synonymous with $T_{comp\_mereology}$*[13].

Proof. Let Δ be the sentence

$$(\forall x, y)\ part^{sumo}(x, y) \leftrightarrow part(x, y))$$

Using Prover9, we can show that $T_{part\_decomp} \models T_{comp\_mereology}$, and $T_{comp\_mereology} \models T_{part\_decomp}$ □

[13] colore.oor.net/ontologies/mereology/comp_mereology.clif.

3.5 Relationship to Classical Mereologies

We have so far evaluated four subtheories of SUMO and proposed revisions to their axiomatizations to address the problem of classes of omitted and unintended models regarding those normally associated with mereology. We now take a closer look at these revised theories.

Proposition 11. *The theory*

$$T_{part} \cup T_{extended\_sum} \cup T_{extended\_product} \cup T_{extended\_decomp}$$

is consistent.

Given that the revised theories are consistent, can we characterize their models? In particular, how is the mereology within SUMO related to the classical mereologies that have been explored by the philosophical and applied ontology communities? Regarding the supplementation principles (22) to (25), respectively named in [21] as *weak company, strong company, supplementation*, and *strong supplementation*, Proposition 12 shows that those principles are not theorems of SUMO.

Proposition 12. *The following sentences are not entailed by* $T_{part} \cup T_{extended\_sum} \cup T_{extended\_product} \cup T_{extended\_decomp}$*:*

$$(\forall x, y) properPart(x, y) \rightarrow \exists z(properPart(z, y) \wedge -(z = x)) \tag{22}$$

$$(\forall x, y) properPart(x, y) \rightarrow (\exists z)(properPart(z, y) \wedge \neg part(z, x)) \tag{23}$$

$$(\forall x, y) properPart(x, y) \rightarrow (\exists z)(Part(z, y) \wedge \neg overlapsSpatially(z, x)) \tag{24}$$

$$(\forall x, y) \neg part(y, x) \rightarrow (\exists z)(Part(z, y) \wedge \neg overlapsSpatially(z, x)) \tag{25}$$

This is closely related to the question of whether or not T_{part} is a module of SUMO. We have already seen that the subtheories of SUMO entail additional mereological theories that are not entailed by T_{part} alone. However, we can see that the addition of these new sentences does form a module of the revised axioms by combining the earlier theorems:

Theorem 6. $T_{part\_sum} \cup T_{part\_prod} \cup T_{part\_decomp}$ *is a module of* $T_{part} \cup T_{extended\_sum} \cup T_{extended\_product} \cup T_{extended\_decomp}$ *that is synonymous with* $T_{strong\_lub\_mereology} \cup T_{prod\_mereology} \cup T_{comp\_mereology}$

3.6 Mereotopology in SUMO

Since mereology can only represent the relation of parts with their respective wholes, predicate *connected* is characterized in SUMO to represent a more general symmetric and reflexive spatial relationship among objects which are not necessarily in a part-whole relation.

Definition 13. *$T_{topology}$ is the subtheory of SUMO consisting of the following axioms:*

$$(\forall x) Object(x) \rightarrow connected(x, x) \tag{26}$$

$$(\forall x, y) connected(x, y) \rightarrow Object(x) \wedge Object(y) \tag{27}$$

$$(\forall x, y) connected(x, y) \rightarrow connected(y, x) \tag{28}$$

The subtheory of SUMO that axiomatizes mereotopology, which we have called $T_{spatial\_relation}$ is intended to characterize the relationship between the notions of mereology and topology. In it, both predicates, *meetsSpatially*, which represents external connection among objects, and *overlapsSpatially*, are declared disjoint specializations of predicate *connected*.

Definition 14. *$T_{spatial\_relation}$ is the subtheory of SUMO which is an extension of $T_{part} \cup T_{topology}$ consisting of the sentences*

$$(\forall x, y)\ meetsSpatially(x, y) \rightarrow connected(x, y) \tag{29}$$

$$(\forall x)\ \neg meetsSpatially(x, x) \tag{30}$$

$$(\forall x, y)\ meetsSpatially(x, y) \rightarrow meetsSpatially(y, x) \tag{31}$$

$$(\forall x, y)\ overlapsSpatially(x, y) \rightarrow connected(x, y) \tag{32}$$

$$(\forall x)\ overlapsSpatially(x, x) \tag{33}$$

$$(\forall x, y)\ overlapsSpatially(x, y) \rightarrow overlapsSpatially(y, x) \tag{34}$$

$$(\forall x, y)\ meetsSpatially(x, y) \rightarrow \neg overlapsSpatially(x, y) \tag{35}$$

$$(\forall x, y) connected(x, y) \rightarrow (meetsSpatially(x, y) \vee overlapsSpatially(x, y)) \tag{36}$$

However, the axiomatization of this theory is already entailed by the following definitional extension of $T_{part} \cup T_{topology}$:

Definition 15. *$T_{mereotop\_def}$ is the definitional extension of $T_{part} \cup T_{topology}$ consisting of the sentences*

$$(\forall x, y) overlapsSpatially(x, y) \leftrightarrow connected(x, y) \wedge (\exists z)\ part(z, x) \wedge part(z, y) \tag{37}$$

$$(\forall x, y) meetsSpatially(x, y) \leftrightarrow connected(x, y) \wedge \neg(\exists z)\ part(z, x) \wedge part(z, y) \tag{38}$$

Proposition 13.

$$T_{mereotop\_def} \models T_{spatial\_relation}$$

We have found that the monotony of relation *connected* with respect to parthood was not characterized in SUMO, which introduces unintended models as the one represented in Fig. 3, where all parts share one point, but only shaded ones result to be connected.

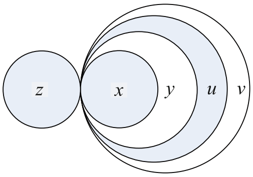


Fig. 3. Model of SUMO where the monotony of relation *connected* with respect to parthood was not characterized. Even though $connected(z,x)$, $part(x,y)$, $part(y,u)$, and$part(u,v)$ hold, $connected(z,y)$ and $connected(z,v)$ do not hold, while $connected(z,u)$ does hold. (Original figure from [16].)

Proposition 14.

$$T_{mereotop\_def} \not\models (\forall x,y) part(x,y) \rightarrow \forall z(connected(z,x) \rightarrow connected(z,y))$$

In order to rule out those unintended models that proposition 14 identifies, we propose the following extension:

Definition 16. *$T_{extend\_mereotop}$ is the theory which extends $T_{part} \cup T_{topology}$ with sentence (39).*

$$(\forall x,y) part(x,y) \rightarrow \forall z(connected(z,x) \rightarrow connected(z,y)) \tag{39}$$

4 DOLCE

The Descriptive Ontology for Linguistic and Cognitive Engineering DOLCE [8, 14] is a freely available upper ontology that is part of the WonderWeb project[14], which is aimed to provide the infrastructure required for a large-scale deployment of ontologies intended to be the foundation for the Semantic Web. DOLCE has a cognitive approach, i.e., it presents the world as it is grasped by humans, based on human knowledge and culture, in opposition to *ontological realism* [9], which intends to present the world as it is, independently of the bias of human perception. The development of DOLCE has followed the principles of the OntoClean methodology [12]. The first version of DOLCE had a representation in Modal Logic, a translation with loss into standard first-order logic, a translation with further loss into OWL, and also an alignment with WordNet [7]. A new version of the fragment of the original ontology that focuses on entities that exist on time, called *temporal particulars*, was presented in [2], called DOLCE-CORE; we will circumscribe our work to the axiomatization of DOLCE-CORE.

At the top of DOLCE-CORE the category of temporal-particulars PT is partitioned into six basic categories: objects O, events E, individual qualities Q, regions R, concepts C, and arbitrary sums AS. Categories ED (*endurant*) and PD (*perdurant*) of DOLCE were, respectively, renamed O (*object*) and

[14] http://wonderweb.semanticweb.org.

E (*event*) in DOLCE-CORE. The axiomatization of mereology in DOLCE-CORE is as follows,[15] where predicate P represents parthood, and (40)–(42) respectively stand for the reflexivity, transitivity, and antisymmetry of relation P. Overlap of parts and mereological sum representing binary fusion of parts are respectively defined in (43) and (44), while (46)–(50) characterize the dissectivity of P across categories, and (51)–(56) close the sum of parts inside each category.

$$(\forall x)P(x,x) \tag{40}$$

$$(\forall x,y)P(x,y) \wedge P(y,z) \rightarrow P(x,z) \tag{41}$$

$$(\forall x,y)P(x,y) \wedge P(y,x) \rightarrow (x=y) \tag{42}$$

$$(\forall x,y)Ov(x,y) \equiv (\exists z)(P(z,x) \wedge P(z,y)) \tag{43}$$

$$(\forall x,y,z)SUM(z,x,y) \equiv (\forall v)Ov(v,z) \leftrightarrow Ov(v,x) \vee Ov(v,y) \tag{44}$$

$$(\forall x,y)\neg P(x,y) \rightarrow (\exists z)P(z,x) \wedge \neg Ov(z,y) \tag{45}$$

$$(\forall x,y)O(y) \wedge P(x,y) \rightarrow O(x) \tag{46}$$

$$(\forall x,y)E(y) \wedge P(x,y) \rightarrow E(x) \tag{47}$$

$$(\forall x,y)T(y) \wedge P(x,y) \rightarrow T(x) \tag{48}$$

$$(\forall x,y)TQ(y) \wedge P(x,y) \rightarrow TQ(x) \tag{49}$$

$$(\forall x,y)C(y) \wedge P(x,y) \rightarrow C(x) \tag{50}$$

$$(\forall x,y,z)O(x) \wedge O(y) \wedge SUM(z,x,y) \rightarrow O(z) \tag{51}$$

$$(\forall x,y,z)E(x) \wedge E(y) \wedge SUM(z,x,y) \rightarrow E(z) \tag{52}$$

$$(\forall x,y,z)T(x) \wedge T(y) \wedge SUM(z,x,y) \rightarrow T(z) \tag{53}$$

$$(\forall x,y,z)TQ(x) \wedge TQ(y) \wedge SUM(z,x,y) \rightarrow TQ(z) \tag{54}$$

$$(\forall x,y,z)C(x) \wedge C(y) \wedge SUM(z,x,y) \rightarrow C(z) \tag{55}$$

$$(\forall x,y,z)AS(x) \wedge AS(y) \wedge SUM(z,x,y) \rightarrow AS(z) \tag{56}$$

Due to the ontological commitment represented by axiom (45), the mereology characterized in DOLCE-CORE is an *extensional mereology*[16] according to [3, 21].

[15] Axioms (48), (49), (53), and (54) are the instantiation of DOLCE higher-order axiom schemas for the subcategories of main categories Q and R which are relevant for our work. A complete version of DOLCE-CORE mereology represented in first-order logic is available at colore.oor.net/ontologies/dolce-core/mereology.in.

[16] It can be proved that in an *extensional mereology* non-atomic entities whose proper parts are the same, are identical, i.e., every entity is exhaustively defined by its parts.

5 Mapping the Mereologies of SUMO and DOLCE

In order to relate SUMO and DOLCE we assume that the changes that we have proposed in Sect. 3 for eliminating unintended models and characterizing missing intended ones have been performed in SUMO. There is no axiomatization in DOLCE-CORE, neither in DOLCE, that corresponds to the notion of topology, therefore our mappings are circumscribed to the axiomatization of mereology in both theories.

By examining the predicates that characterize the participation of objects in events in both ontologies, and also by the type of relation that the main categories of SUMO and DOLCE-CORE have with time and space, we have built the translation definitions of Table 1 for the main these categories.

Table 1. Mapping of SUMO and DOLCE main categories.

$(\forall x) Object(x) \leftrightarrow O(x)$	(57)
$(\forall x) Process(x) \leftrightarrow E(x)$	(58)
$(\forall x) TimeInterval(x) \leftrightarrow T(x)$	(59)
$(\forall x) Region(x) \leftrightarrow S(x)$	(60)

Table 2. Translation definitions for $T_{dolce\_part\_t}$ into $T_{time\_mereology}$.

$(\forall x) T(x) \leftrightarrow TimeInterval(x)$	(61)
$(\forall x, y) P(x, y) \leftrightarrow temporalPart(x, y)$	(62)
$(\forall x, y) Ov(x, y) \leftrightarrow overlapsTemporally(x, y)))$	(63)

Table 3. Translation definitions for $T_{time_{m}ereology}$ into $T_{dolce\_part\_t}$.

$(\forall x) TimeInterval(x) \leftrightarrow T(x)$	(64)
$(\forall x, y) temporalPart(x, y) \leftrightarrow P(x, y) \wedge T(x) \wedge T(y)$	(65)
$(\forall x, y) overlapsTemporally(x, y) \leftrightarrow Ov(x, y) \wedge T(x) \wedge T(y)$	(66)

5.1 Mapping Time

The subtheory T_{sumo_time}, whose modular structure is shown in Fig. 5, characterizes the axiomatization of time in SUMO. This theory, which was verified in [20], includes 3 submodules[17] $T_{sumo\_ordered\_timepoints}$, $T_{sumo\_timeintervals}$, and $T_{time\_mereology}$, such that each module is a conservative extension of each connected subtheory below it in Fig. 5. These 3 subtheories respectively characterize a linear ordering between instants of time, a part-whole relation among intervals of time, and an account of Allen's interval relations *starts, finishes, during, earlier*, and *meetsTemporally* [13]. Finally, T_{sumo_time} characterizes a part-whole relationship that includes intervals and instants of time.

On the other hand, T_{dolce_part} characterizes parthood by unique predicate P across every category, including T. By means of the following theorems we can characterize the relationship that exists among $T_{dolce\_part\_t}$ and T_{sumo_time}.

[17] Available at colore.oor.net/ontologies/sumo/modules.

Theorem 7. *Let $T_{dolce\_part\_t}$ be the subtheory of T_{dolce_part} with signature $\{T, part\}$, and Let $T_{time\_mereology}$ be the theory given by axioms (67)–(72). Then, $T_{time\_mereology}$ is synonymous with $T_{dolce\_part\_T}$.*

$$(\forall x) TimeInterval(x) \rightarrow temporalPart(x, x). \tag{67}$$

$$(\forall x, y) temporalPart(x, y) \wedge temporalPart(y, x) \rightarrow (x = y). \tag{68}$$

$$(\forall x, y, z) temporalPart(x, y) \wedge temporalPart(y, z) \rightarrow temporalPart(x, z). \tag{69}$$

$$(\forall x, y) overlapsTemporally(x, y) \rightarrow TimeInterval(x) \wedge TimeInterval(y) \tag{70}$$

$$(\forall x) TimeInterval(x) \rightarrow overlapsTemporally(x, x)). \tag{71}$$

$$\begin{split}(\forall x, y) TimeInterval(x) \wedge TimeInterval(y) \rightarrow (overlapsTemporally(x, y) \leftrightarrow \\ ((\exists z)(TimeInterval(z) \wedge temporalPart(z, x) \wedge temporalPart(z, y))))\end{split} \tag{72}$$

Proof. Let Δ be the set of translations shown in Table 2, and Υ the set of translations shown in Table 3. Using Prover9 we have shown that $T_{time\_mereology} \cup \Delta \models T_{dolce\_part\_T}$, and $T_{dolce\_part\_T} \cup \Upsilon \models T_{time\_mereology}$. □

5.2 Mapping Events

Regarding the representation of events in SUMO and DOLCE, by means of the following definition and theorem we classify the relationship that their respective part-whole axiomatizations have as *synonymy*.

Table 4. Translation definitions for $T_{dolce\_part\_E}$ into $T_{sumo\_subprocess}$.

$(\forall x) E(x) \leftrightarrow Process(x)$	(73)
$(\forall x, y) P(x, y) \leftrightarrow subProcess(x, y)$	(74)
$(\forall x, y) Ov(x, y) \leftrightarrow (\exists z)(subProcess(z, x) \wedge subProcess(z, y))$	(75)

Table 5. Translation definitions for $T_{sumo\_subprocess}$ into $T_{dolce\_part\_E}$.

$(\forall x) Process(x) \leftrightarrow E(x)$	(76)
$(\forall x, y) subProcess(x, y) \leftrightarrow E(x) \wedge E(y) \wedge P(x, y)$	(77)

Definition 17. *$T_{sumo\_subprocess}$ is the theory given by the axioms:*

$$(\forall x, y) subProcess(x, y) \to Process(x) \wedge Process(y) \tag{78}$$

$$(\forall x) Process(x) \to subProcess(x, x) \tag{79}$$

$$(\forall x, y) subProcess(x, y) \wedge subProcess(y, z) \to subProcess(x, z) \tag{80}$$

$$(\forall x, y) subProcess(x, y) \wedge subProcess(y, x) \to (x = y) \tag{81}$$

Theorem 8. *Let $T_{dolce\_part\_E}$ be the theory given by axioms (40)–(42) and (47). $T_{sumo\_subprocess}$ is synonymous with $T_{dolce\_part\_E}$.*

Proof. Let Δ be the set of translations shown in Table 4 and Γ the set of translations shown in Table 5. Using Prover9 we have demonstrated that $T_{sumo\_subprocess} \cup \Delta \models T_{dolce\_part\_E}$ and $T_{dolce\_part\_E} \cup \Gamma \models T_{sumo\_subprocess}$. □

5.3 Mapping Objects

Regarding the representation of objects in SUMO and DOLCE-CORE, by means of the following theorem we classify the relationship among their respective part-whole axiomatizations as *synonymy*.

Definition 18. SUMO PART *is the theory given by axioms (1)–(7), and* DOLCE PART-T *is the theory given by axioms (40)–(43) and (46).*

Theorem 9. *Let T_{sumo_part} be the theory given by axioms (1)–(7), and $T_{dolce\_part\_O}$ the theory given by axioms (40)–(43) and (46). Then, T_{sumo_part} is synonymous with $T_{dolce\_part\_O}$.*

Proof. Let us call Δ to the set of translations shown in Table 6, and Π the set of translations shown in Table 7. Using Prover9 we have shown that $T_{sumo\_part} \cup \Delta \models T_{dolce\_part\_O}$ and $T_{dolce\_part\_O} \cup \Pi \models T_{sumo\_part}$. □

Table 6. Translations DOLCE PART-O into SUMO PART.

$(\forall x, y) P(x, y) \leftrightarrow part(x, y))$	(82)
$(\forall x, y) Ov(x, y) \leftrightarrow overlapsSpatially(x, y))$	(83)

Table 7. Translations SUMO PART into DOLCE PART-O.

$(\forall x, y) part(x, y) \leftrightarrow O(x) \wedge O(y) \wedge P(x, y)$	(84)
$(\forall x, y) properPart(x, y) \leftrightarrow O(x) \wedge O(y) \wedge P(x, y) \wedge \neg P(y, x)$	(85)
$(\forall x, y) overlapsSpatially(x, y) \leftrightarrow O(x) \wedge O(y) \wedge Ov(x, y))$	(86)
$(\forall x, y) overlapsPartially(x, y) \leftrightarrow Ov(x, y) \wedge \neg P(x, y))) \wedge \neg P(y, x))))$	(87)

5.4 Mapping Mereologies with Sums

The theories T_{dolce_sum} in DOLCE and $T_{extended\_sum}$ for SUMO are both intended to axiomatize the intuitions regarding the fusion of parts. The key question is now whether or not they actually axiomatize the same class of intended models.

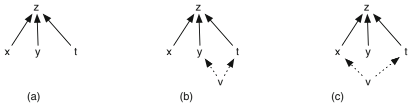


Fig. 4. Objects x, y, z, t, which do not hold $SUM(z,x,y)$ but hold $MereologicalSumFn(x,y) = z$. Arrows represent relation *part* of theory $T_{extended\_sum}$, and relation P of theory T_{dolce_sum}. (Original figure from [16]).

Table 8. Translations DOLCE SUM into SUMO SUM.

$$(\forall x,y,z)SUM(z,x,y) \leftrightarrow Object(x) \wedge Object(y) \wedge (MereologicalSumFn(x,y) = z) \tag{88}$$

The axiomatization of $T_{extended\_sum}$ from SUMO is weaker than the axiomatization of T_{dolce_sum} in DOLCE. In fact, let us consider objects x, y, z, t of Fig. 4, such that $properPart(x,z)$, $properPart(y,z)$, and $properPart(t,z)$ hold, while none of $overlapsSpatially(x,y)$, $overlapsSpatially(t,y)$, or $overlapsSpatially(x,t)$ hold. In parts (a), (b), and (c) of the bottom of Fig. 4 parthood is indicated with arrows from the part to the whole. According to the characterization of mereological sum in $T_{extended\_sum}$, we must have $(MereologicalSumFn(x,y) = z)$. However, parts (b) and (c) of Fig. 4 depict alternative additional conditions that the characterization of mereological sum in T_{dolce_sum} must satisfy. In DOLCE, any other object t which overlaps with the sum z must also overlap with at least one of the addends x or y, which is a condition indicated by dotted lines. Because neither $overlapsSpatially(x,t)$ nor $overlapsSpatially(y,t)$ hold, then $SUM(z,x,y)$ does not hold in DOLCE. The following theorem formalizes our claim.

Theorem 10. *$T_{extended\_sum}$ cannot interpret T_{dolce_sum}.*

Proof. Let us call Δ to the translations shown in Table 6, and Π to the translation shown in Table 8, and let T_1 be the theory that results from adding sentence (89) to theory $T_{extended\_sum}$. Using Mace4, we have built a model of $T_1 \cup \Delta \cup \Pi$ (see footnote 17).

$$(\exists x,y,z)SUM(z,x,y) \wedge \neg(\forall w)(Ov(w,z) \leftrightarrow Ov(w,x) \vee Ov(w,y)) \tag{89}$$

□

In order to translate the symbol $MereologicalSumFn$ of theory $T_{extended\_sum}$ into the language of DOLCE-CORE, we have represented the graph[18] of function *MereologicalSumFn* by means of predicate *MSum*, as shown in Table 9.

Theorem 11. *T_{dolce_sum} cannot interpret $T_{extended\_sum}$.*

Proof. Let us call Δ to the translations shown in Table 1, Π to the translations in Table 7, and Υ to the translation in Table 10, and let T_1 be the theory that results from adding sentence (90) to T_{dolce_sum}. Using Mace4, we have built a model of $T_1 \cup \Delta \cup \Pi \cup \Upsilon$ (see footnote 18).

$$(\exists x, y)Object(x) \wedge Object(y) \wedge (\forall z)(\neg Object(z) \vee \neg MSum(z, x, y)) \tag{90}$$

□

Table 9. Characterization of predicate MSum in SUMO.

$$(\forall x, y, z)MSum(z, x, y) \rightarrow (\forall w)(part(z, w) \leftrightarrow (part(x, w) \wedge part(y, w)))) \tag{91}$$

$$(\forall x, y, z)MSum(z, x, y) \rightarrow Object(x) \wedge Object(y) \wedge Object(z) \tag{92}$$

$$(\forall x, y)Object(x) \wedge Object(y) \rightarrow \exists z(Object(z) \wedge MSum(z, x, y)) \tag{93}$$

$$(\forall x, y, z, t)MSum(z, x, y) \wedge MSum(t, x, y) \rightarrow (z = t) \tag{94}$$

Table 10. Translation of $T_{extended\_sum}$ into T_{dolce_sum}.

$$(\forall x, y, z)MSum(z, x, y) \leftrightarrow Object(x) \wedge Object(y) \wedge SUM(z, x, y) \tag{95}$$

Figure 5 shows conservative extensions by means of thin black arrows and relative interpretations (mappings), by thick grey arrows from interpreted to interpreting theories. Because every theorem of a theory is also a theorem of its conservative extensions, each conservative extension is capable of interpreting every theory that the modules that it extends interpret. In particular, the subtheory T_{dolce_part}, shown in Fig. 5, is the theory resulting from the union of $T_{dolce\_part\_T}$, $T_{dolce\_part\_E}$, and $T_{dolce\_part_O}$, plus axioms (49), (50), while the subtheory *DOLCE EXTENSIONAL MEREOLOGY* is the union of T_{dolce_part}, T_{dolce_sum}, and axioms

[18] A n-ary function f from A^n to B is representable by a relation ϱ with arity (n+1), called *the graph of f*, such that:
(a) Every tuple of ϱ is a tuple $\langle \bar{x}, f(\bar{x}) \rangle$ with $\bar{x} \in A^n$ and $f(\bar{x}) \in range(f)$.
(b) If $f(\bar{x}) = b$ and $f(\bar{z}) = c$, then $b = c$.

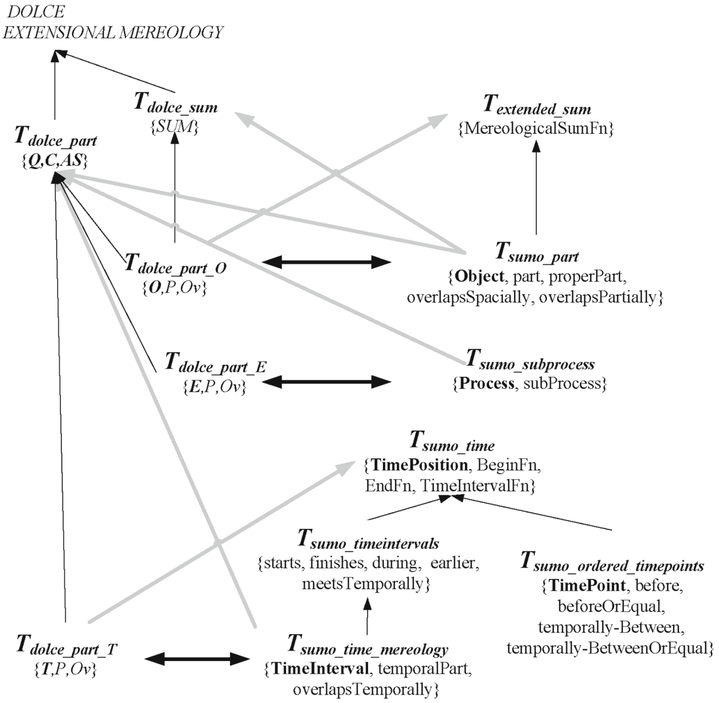

DOLCE
EXTENSIONAL MEREOLOGY
T_{dolce_sum}
{SUM}
$T_{extended\_sum}$
{MereologicalSumFn}
T_{dolce_part}
{Q,C,AS}
$T_{dolce\_part\_O}$
{O,P,Ov}
T_{sumo_part}
{Object, part, properPart, overlapsSpacially, overlapsPartially}
$T_{dolce\_part\_E}$
{E,P,Ov}
$T_{sumo\_subprocess}$
{Process, subProcess}
T_{sumo_time}
{TimePosition, BeginFn, EndFn, TimeIntervalFn}
$T_{sumo\_timeintervals}$
{starts, finishes, during, earlier, meetsTemporally}
$T_{sumo\_ordered\_timepoints}$
{TimePoint, before, beforeOrEqual, temporally-Between, temporally-BetweenOrEqual}
$T_{dolce\_part\_T}$
{T,P,Ov}
$T_{sumo\_time\_mereology}$
{TimeInterval, temporalPart, overlapsTemporally}

Fig. 5. Mappings between modules of DOLCE-CORE and SUMO (extracted from [16]). Black thin arrows point to conservative extensions, thick grey arrows are directed from interpreted theories to interpreting theories, and thick black arrows connect synonymous theories.

(45), (52), (53), (54), (55), and (56). As indicated by oriented grey arrows, the axiomatization of part-whole relations in categories *Object*, *Process*, and *TimeInterval* of SUMO are mappable to DOLCE minimal axiomatization of mereology represented by the subtheory T_{dolce_part}. Although not represented in Fig. 5, it holds that because *DOLCE EXTENSIONAL MEREOLOGY* extends T_{dolce_part}, it also interprets T_{sumo_part}, $T_{sumo\_subprocess}$, and $T_{time\_mereology}$. In turn, T_{sumo_sum} interprets $T_{dolce\_part\_O}$. The strongest subtheories of SUMO and DOLCE-CORE that are synonymous, and therefore have equivalent models, are the pairs indicated by double black arrows, i.e, $T_{dolce\_part\_O}$ with T_{sumo_part}, $T_{dolce\_part\_E}$ with $T_{sumo\_subprocess}$, and $T_{dolce\_part\_T}$ with $T_{time\_mereology}$.

6 Conclusions

Since the conceptual coverage of upper ontologies needs to be broad enough to cover the underpinnings of domain ontologies, we find a modules of upper ontologies for notions about time, space, objects, and processes. In this paper, we have focussed on how two upper ontologies (SUMO and DOLCE) axiomatize mereotopologies, that is, the notions of parthood (mereology) and connection (topology). By showing how different subtheories of SUMO and DOLCE are logically synonymous with different theories of lattices, we have identified unintended and omitted models of the original axiomatization of SUMO. There are additional subtheories of SUMO that capture additional spatial concepts such as containment, holes, orientation, and betweenness. Future work will provide a verification of these modules, thereby provided a firmer foundation for spatial reasoning.

References

1. Benzmüller, C., Pease, A.: Higher-order aspects and context in SUMO. J. Web Sem. **12**, 104–117 (2012)
2. Borgo, S., Masolo, C.: Foundational choices in DOLCE. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies. IHIS, pp. 361–381. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_16
3. Casati, R., Varzi, A.C.: Parts and Places: The Structures of Spatial Representation. A Bradford Book. MIT Press, Cambridge (1999)
4. Davey, B.A., Priestley, H.A.: Introduction to Lattices and Order. Cambridge University Press, Cambridge (2002). Cambridge mathematical text books
5. Enderton, H.B.: A Mathematical Introduction to Logic. Academic Press, Cambridge (1972)
6. Euzenat, J., Shvaiko, P.: Ontology Matching, 2nd edn. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38721-0
7. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A.: Sweetening WORDNET with DOLCE. AI Mag. **24**(3), 13–24 (2003)
8. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 166–181. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45810-7_18
9. Grenon, P., Smith, B.: SNAP and SPAN: towards dynamic spatial ontology. Spat. Cogn. Comput. **4**(1), 69–103 (2004)
10. Grüninger, M., Hahmann, T., Hashemi, A., Ong, D.: Ontology verification with repositories. In: Proceedings of the Sixth International Conference Formal Ontology in Information Systems, FOIS 2010, Toronto, Canada, 11–14 May 2010, pp. 317–330 (2010)
11. Guarino, N.: Formal ontology and information systems. In: Formal Ontology in Information Systems - Proceedings of FOIS 1998, Trento, Italy, 6–8 June 1998, pp. 3–15. IOS Press, Amsterdam, pp. 3–15 (1998)
12. Guarino, N., Welty, C.: Evaluating ontological decisions with ontoclean. Commun. ACM **45**(2), 61–65 (2002)
13. Hayes, P.: Catalog of temporal theories. Technical report UIUC-BI-AI-96-01. University of Illinois Urbana-Champagne (1996)
14. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: WonderWeb deliverable D18 ontology library (final). Technical report, IST Project 2001–33052 WonderWeb: Ontology Infrastructure for the Semantic Web (2003)
15. McCune. W.: Prover9 and Mace4 (2005–2010). http://www.cs.unm.edu/mccune/prover9/

16. Silva Muñoz, L., Grüninger, M.: Verifying and mapping the mereotopology of upper-level ontologies. In: Knowledge Engineering and Ontology Design, KEOD 2016, Porto, Portugal, 9–11 November 2016, pp. 31–42 (2016)
17. Niles, I., Pease, A.: Linking lexicons and ontologies: mapping WordNet to the suggested upper merged ontology. In: Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 2003), Las Vegas, Nevada (2003)
18. Niles, I., Pease, A.: Towards a standard upper ontology. In FOIS 2001: Proceedings of the international conference on Formal Ontology in Information Systems, pp. 2–9. ACM, New York (2001)
19. Pearce, D., Valverde, A.: Synonymous theories and knowledge representations in answer set programming. J. Comput. Syst. Sci. **78**(1), 86–104 (2012)
20. Silva Muñoz, L., Grüninger, M.: Mapping and verification of the time ontology in SUMO. In: Formal Ontology in Information Systems - Proceedings of the 9th International Conference, FOIS, 6–9 July 2016, Annecy, France (2016)
21. Varzi, A.C.: Spatial reasoning and ontology: parts, wholes, and locations. In: Aiello, M., Pratt-Hartmann, I., Van Benthem, J. (eds.) Handbook of Spatial Logicspages, pp. 945–1038. Springer, Dordrecht (2007). https://doi.org/10.1007/978-1-4020-5587-4_15

SerVCS: Serialization Agnostic Ontology Development in Distributed Settings

Lavdim Halilaj[1,2](✉), Irlán Grangel-González[1,2], Maria-Esther Vidal[2,3,4], Steffen Lohmann[2], and Sören Auer[3,4]

[1] Smart Data Analytics (SDA), University of Bonn, Bonn, Germany
{halilaj,grangel}@cs.uni-bonn.de

[2] Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Sankt Augustin, Germany
steffen.lohmann@iais.fraunhofer.de

[3] TIB Leibniz Information Center for Science and Technology, Hannover, Germany
{vidal,auer}@l3s.de

[4] L3S Research Center, University of Hannover, Hannover, Germany

Abstract. The development of domain-specific ontologies requires joint efforts among different groups of stakeholders, such as knowledge engineers and domain experts. During the development processes, ontology changes need to be tracked and propagated across developers. Version Control Systems (VCSs) collect metadata describing changes and allow for the synchronization of different versions of the same ontology. Commonly, VCSs follow *optimistic* approaches to enable the *concurrent* modification of ontology artifacts, as well as conflict detection and resolution. For conflict detection, VCSs usually apply techniques where files are compared line by line. However, ontology changes can be serialized in different ways during the development process. As a consequence, existing VCSs may detect a large number of *false-positive* conflicts, i.e., conflicts that do not result from ontology changes but from the fact that two ontology versions are differently serialized. We developed *SerVCS* in order to enhance VCSs to cope with different serializations of the same ontology, following the principle of *prevention is better than cure*. *SerVCS* resorts on unique ontology serializations and minimizes the number of false-positive conflicts. It is implemented on top of Git, utilizing tools such as Rapper and RDF-toolkit for syntax validation and unique serialization, respectively. We conducted an empirical evaluation to determine the conflict detection accuracy of *SerVCS* whenever simultaneous changes to an ontology are performed using different ontology editors. Experimental results suggest that *SerVCS* allows VCSs to conduct more effective synchronization processes by preventing false-positive conflicts.

1 Introduction

During ontology development, the number, structure, and terminology of the modeled concepts and relations is subject to continuous change. The development process usually requires significant efforts and knowledge, and the participation of different stakeholders who are geographically distributed [1]. One of

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 213–232, 2019.
https://doi.org/10.1007/978-3-319-99701-8_10

the main challenges for the involved ontology engineers is to work collaboratively on a shared objective in a harmonic and efficient way, while avoiding misunderstandings, uncertainty, and ambiguity [2]. It is crucial in this process, tracking and propagating ontology changes to all contributors, and users should be able to synchronize changes with their work. Thus, supporting change management is indispensable for successful ontology development in distributed settings.

A *Version Control System* (VCS) assists users in working collaboratively on shared artifacts, and helps to prevent them from overwriting changes made by others. Basically, mechanisms to avoid change overwriting can be classified in *pessimistic* and *optimistic* approaches [3]. The first ones are based on the *lock-modify-unlock* paradigm, which implies that modifications to an artifact are permitted only for one user at a time. The latter ones are based on the *copy-modify-merge* paradigm, where users work on personal copies, each reflecting the remote repository at a certain time. After the work is completed, the local changes are merged into the remote repository by an *update* command, comprising the phases *comparison*, *conflict detection*, *conflict resolution*, and *merge*.

Different techniques, such as line-, tree-, and graph-based ones, can be employed to compare two versions of the same artifact [4]. The line-based technique, which achieved wide applicability, compares artifacts line by line, with each line being treated as a single unit. This technique is also known as *textual* or *line-based comparison* [3]. Examples of VCSs that use the line-based approach are Subversion, CVS, Mercurial, and Git. Line-based comparisons are applicable on any kind of text artifact, as they do not consider syntactical information [4]. Accordingly, line-based approaches also neglect syntactical information of ontologies, which are commonly represented in some text-based OWL serialization.

Challenges arise when two ontology developers modify same artifacts on their personal working copies in parallel. Changes might contradict each other, for instance, developers may both edit the name of an ontology concept simultaneously. Such parallel and controversial modifications can result in conflicts during the merging of two ontology versions. In general, a conflict is defined as "a set of contradicting changes where at least one operation applied by the first developer does not commute with at least one operation applied by the second developer" [4]. Conflicts can be detected by identifying changed units (i.e., added, updated, deleted) in parallel. Conflict resolution can be done automatically or may require users to manually fix them by resolving the conflictual changes.

From the ontology development point of view, the situation is exacerbated when different ontology editors are used during the development process. This is due to the fact that these editors often produce different serializations of the same ontology, i.e., the ontology concepts are grouped and sorted differently in the files generated by the editors.[1] As a result, the ability of VCSs to detect the actual changes in ontologies is lowered, since they find a number of conflicts that are actually not given but are a result of different serializations of the ontology

[1] With "different serializations", we refer to two different ontology files that represent the same ontology using the same syntax (RDF/XML, Turtle, Manchester, etc.) but use a different structure to list and group the ontology concepts.

file. In order to increase the accuracy of conflict detection in VCSs, the problem of different groupings and orderings must be tackled.

In this paper, we present *SerVCS*, a generic approach for the realization of optimistic and tool-independent ontology development on the basis of Version Control Systems. As a result, VCSs become *editor agnostic*, i.e., capable to detect actual changes and automatically resolve conflicts using the *built-in* merging algorithms. We implemented and applied the *SerVCS* approach on the basis of the widely used Git VCS. In addition, we developed a middleware service to generate a unique serialization of ontologies before they are pushed to the remote repository. The unique serialization ensures that ontologies have always the same serialization in the remote repository, regardless of the used ontology editor. Thus, we avoid incompatibility problems with regard to wrongly detected conflicts resulting from the use of different ontology editors, and assist ontology developers to collaborate more efficiently in distributed environments.

This paper is an extension of our previous work [5], where we presented the initial idea and approach to prevent false-positive conflicts in distributed ontology development, when different editors and serializations are used. The novel contributions presented in this paper are summarized as follows:

- an extensive analysis of the state-of-the-art and the positioning of our work with respect to existing approaches;
- a more flexible architecture, allowing for the integration of other tools that generate unique serializations using different sorting criteria; and
- a comprehensive empirical evaluation that demonstrates the impact of ontology size and sorting criteria on the effectiveness of our approach.

The remainder of this paper is organized as follows: After a comparison of our work with the state-of-the-art in Sect. 2, a motivating scenario is presented in Sect. 3. In Sect. 4, we describe the *SerVCS* approach, which is complemented by a description of its implementation in Sect. 5. The approach is evaluated against concrete cases in Sect. 6, before the paper is concluded with an outlook on possible extensions in Sect. 7.

2 Related Work

Synchronization of changes among different versions of the same artifact has attracted the attention of researches for several years. For example, enhancing VCSs with additional information related to the semantics of software code and re-factoring with the objective of improving the merging process has been proposed in a number of works [6–8]. Asenov et al. [9,10] and Protzenko et al. [10] propose to add *unique IDs* to source code elements in order to achieve a more precise detection of conflicts. Brun et al. [11] present an approach called “speculative analysis”, identifying the possible existence of conflicts in a continuous and precise way. Furthermore, they identify several classes of conflicts and provide detailed instructions how to address them.

However, these approaches are focused on source code of software artifacts where order line is important. Although, line order can also be important in our case, semantics encoded in the ontology is not necessarily affected by line order.

2.1 Version Management for Ontologies

Approaches that focus on providing version management for ontologies in collaborative development processes are discussed. An ontology for unique identification of changes between two RDF graphs is presented by Lee et al. [12]. To recognize these changes (or deltas), a pretty-printed version of RDF graphs is utilized. The authors distinguish two types of deltas that can be applied as patches to RDF graphs. First, *weak* deltas, which are directly applied to the graph from where they are computed. Second, *strong* deltas, which specify the changes independently of the context. In contrast to *SerVCS*, this approach focuses on the semantic representation of changes and its application to RDF graphs.

Vöelkel et al. [13] present *SemVersion*, an RDF-based system for ontology versioning. The approach is based on the two core components *data management* and *versioning functionality*. The first is responsible for the storage and retrieval of data chunks. The second deals with specific features of the ontology language, such as structural and semantic differences. To find semantic differences between two versions, e.g., whether a statement has been added or removed, SemVersion employs a simplified heuristic method for conflict detection.

Cassidy et al. [14] propose an approach for realizing a distributed version control system for RDF stores. The approach is based on a semi-formal theory of patches to allow for the manipulation of so-called RDF-patches with the objective of facilitating the revert and merge operations. An implementation of the approach has been realized on top of the VCS Darcs http://darcs.net/ which enables linguistic annotation on RDF stores among different users.

A holistic approach for collaborative ontology development based on ontology change management is described by Palma et al. [1]. The approach comprises different strategies and techniques to realize collaborative processes in inter-organizational settings, such as centralized, decentralized, and hybrid ones.

Edwards [15] proposes techniques for managing *high-level* application-defined conflicts. Consequently, the introduced mechanisms should be able to handle conflict resolutions. Further, certain types of conflicts can be tolerated and others forbidden according to the specified application requirements. Furthermore, a multi-editor environment for collaborative ontology editing is presented by Noy et al. [16]. The proposed framework is able to control and maintain ontologies, as well as support users throughout the whole ontology development process.

The majority of the above mentioned approaches [1,13,15,16] rely on their own version control mechanisms tailored for ontology development. Other approaches [12,14] utilize a semi-formal theory of patches to find deltas among versions, thus enabling to revert or merge specific versions. Contrary, *SerVCS* targets and empowers generic VCSs; thus, conflicts are more precisely detected.

2.2 Version Management for Model-Based Development

We describe works whose main focus is on overcoming the problem of wrongly indicated conflicts in the field of *model-based development*, where *models* are the main artifact. *SMoVer*, a semantically enhanced Version Control System for models, is proposed by Altmanninger et al. [17,18]. Using the semantic view concept to explain aspects of a modeling language, a better conflict detection can be achieved and the reason of conflicts can be more easily determined.

Brosch [19] suggests using a model checker for detecting semantic merge conflicts of an evolving UML sequence diagram. When an automatic merge is not possible due to conflicting changes, additional redundant information essential for the models is used to determine invalid solutions. By using this technique, it is possible to assert concrete modifications realized in a sequence diagram.

A related technique uses domain-specific metamodels describing syntactic and semantic conflicts associated with their resolution [20]. This technique allows for the identification of different conflict patterns that occur during the modeling phase, which are frequently ignored by common structure-based algorithms.

Krusche et al. [21] tackle real-time model synchronization, and propose EMF-Store, a peer-to-peer based solution for real-time synchronization of changes on model instances with all collaborators of a session. EMFStore borrows concepts of VCS's to synchronize models in real-time but does not include a VCS itself.

Zhang et al. [22] investigate composite-level conflict detection in UML Model Versioning, and propose a two-fold approach: In the preprocessing stage, redundant operations are removed from the originally recorded operation lists. During the conflict detection stage, a fragmentation procedure is performed to collect only potentially conflicted operations into the same fragment. Finally, a pattern-matching strategy is followed to solve the conflict detection problem.

In contrast to the aforementioned works, *SerVCS* focuses on ontologies as main artifacts. Moreover, it utilizes functionalities of existing VCSs to merge versions of the same ontology created by different editors.

3 Motivating Example

As a motivating example, we consider two users working together in developing an ontology for a specific domain. In order to ease collaboration and maintain different versions of the developed ontology that result from changes, they decide to use Git. They proceed by setting up the working environment and creating an initial ontology repository which contains several files. Together, the users define the ontology structure with the most fundamental concepts and upload the ontology file F to the *remote repository.* After that, they decide to proceed with their tasks by separately working on their local machines.

The users start synchronizing their local working copies with the remote repository, as illustrated in *Scene 1* of Fig. 1. *Scene 2* depicts simultaneous changes performed on different copies of the same ontology file, such as adding new concepts, modifying existing ones, or deleting concepts. For realizing this

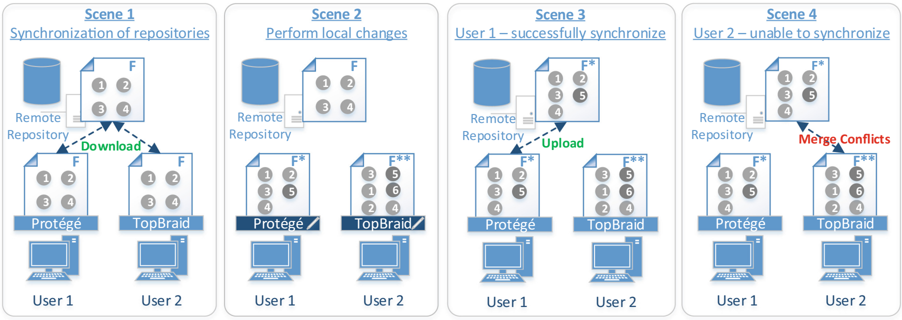


Fig. 1. Motivating example [5]. A distributed environment illustrating an ontology development process. Different ontology editors, e.g., Editors X and Y, are used for defining ontology F by Users 1 and 2. F* and F** represent local versions of F. If F* is uploaded first, changes in F* can be synchronized. Changes in F** cannot be merged whenever F* and F** serializations are different.

task, different ontology editors are used. In our case, *User 1* works with *Desktop Protégé* http://protege.stanford.edu, whereas *User 2* prefers to edit the ontology with *TopBraid Composer* http://www.topquadrant.com/composer/.

After finishing the task, *User 1* uploads her personal working copy (*F**) to the *remote repository*, as shown in *Scene 3*. Next, *User 2* completes his task and starts uploading the changes he made on his local copy to the *remote repository*. While trying to trigger this action, he receives a rejection message from the VCS, listing all changes which result in conflicts, as depicted in *Scene 4*. These conflicts need to be resolved in order for the VCS to allow the user to successfully upload his version (*F***) to the *remote repository*. *User 2* starts resolving the conflicts manually by comparing his version of the ontology with the one of *User 1* that has already been uploaded to the remote repository.

Since the users are working with different ontology editors that use each its own serialization when saving the ontology file, the files are differently organized. For instance, while the concepts in one of the files are grouped into categories, such as *Classes* and *Properties*, they are ordered alphabetically in the other case, without any grouping. Consequently, the information about actual changes, i.e., concrete changes on the ontology performed by each user, can no longer be detected by the line-based comparison of the VCS, but a huge number of conflicts result that are due to the different organization of the ontology files. This prevents *User 2* from merging his changes, and his version of the ontology cannot be uploaded to the remote repository.

This scenario illustrates that, despite the various benefits provided by a VCS for collaborative ontology development, it has not been possible so far to effectively use a VCS in cases where different editors and ontology serializations are used. This is changed with the *SerVCS* approach we present in this paper.

4 Approach

Basic terminology and provide a formal description of the *SerVCS* approach are defined. An ontology is represented in an RDF document, which A is formally defined as $A \subset (\mathbf{I} \cup \mathbf{B}) \times \mathbf{I} \times (\mathbf{I} \cup \mathbf{B} \cup \mathbf{L})$, where $\mathbf{I}$, $\mathbf{B}$, and $\mathbf{L}$ correspond to sets of *IRI*s, *blank nodes*, and *literals* (typed and untyped), respectively [23].

Definition 1 (Changeset). *Given two RDF documents A and A^*, a changeset of A^* with respect to A is defined as follows:*

$$ChangeSet(A^*/A) = (\delta^+(A^*/A), \delta^-(A^*/A), <),\ where$$

- $\delta^+(A^*/A) = \{t \mid t \in A^* \wedge t \notin A\}$,
- $\delta^-(A^*/A) = \{t \mid t \in A \wedge t \notin A^*\}$, *and*
- $<$ *is a partial order between the RDF triples in* $\delta^+(A^*/A) \cup \delta^-(A^*/A)$.

Example 1. Consider two RDF documents $A = \{t_1, t_2, t_3\}$ and $A^* = \{t_1, t_2, t_4\}$ such that A^* is a new version of A where the RDF triple t_4 was added and the triple t_3 was deleted. Then, the changeset of A with respect to A^*, $ChangeSet(A^*/A)$, is as follows:

- $\delta^+(A^*/A) = \{t_4\}$,
- $\delta^-(A^*/A) = \{t_3\}$, and
- $\leq \{(t_4, t_3)\}$.

Definition 2 (Syntactic Conflicts). *Given two RDF documents A and A^*, and the changeset of A^* with respect to A, $ChangeSet(A^*/A) = (\delta^+(A^*/A), \delta^-(A^*/A), <)$, there is a syntactical conflict between A and A^* iff there are RDF triples t_i and t_j such that:*

- $t_i \in \delta^-(A^*/A)$,
- $t_j \in \delta^+(A^*/A)$,
- $(t_i, t_j) \in <$, *and*
- $t_i = (s, p, o_i)$, $t_j = (s, p, o_j)$, *and* $o_i \neq o_j$.

Example 2. Consider two RDF documents A and A^* with triples $t_3 =$ (`:Train, rdfs:label, "Trai"@en`) and $t_4 =$ (`:Train, rdfs:label, "Trainn"@en`). Since the object value of the property `rdfs:label` of the subject `:Train` has been changed, there is a syntactic conflict between the RDF documents A and A^*.

Definition 3 (RDF Document Serialization). *Given an RDF document A and an ordering criteria η, a* serialization *of A according to η, $\Gamma(A, \eta)$ corresponds to an ordering of the triples in A according to η:*

$$\Gamma(A, \eta) = <t_1, t_2, \ldots, t_n>$$

Example 3. Suppose three RDF triples t_1, t_2, and t_3 are defined as follows in an RDF document A: t_1=(`:Car, rdfs:label, "Car"@en`), $t_2 =$ (`:Truck, rdfs:label, "Truck"@en`), and $t_3 =$ (`:Bus, rdfs:label, "Bus"@en`), respectively. A serialization $\Gamma(A, \eta)$ of A listing the triples by their labels in alphabetical order η would be:

$$\Gamma(A, \eta) = <t_3, t_1, t_2>

Definition 4 (False-Positive Conflicts). *Given two RDF documents A and A^* such that F_1 and F_2 are serializations of A and A^* according to some ordering criteria η_1 and η_2, respectively. There is a* false-positive conflict *between F_1 and F_2, iff there exist η ordering criteria such that:*

$$\Gamma(A, \eta) = \Gamma(A^*, \eta) \text{ and } F_1 \neq F_2$$

Example 4. Consider serializations $F_1 = <t_1, t_3, t_2>$ and $F_2 = <t_2, t_1, t_3>$ both representing two identical RDF documents $A = A^*$, respectively, such that $A = \{t_1, t_2, t_3\}$. Then, there are three false-positive conflicts between F_1 and F_2, because there exist ordering criteria η, $\Gamma(A, \eta) = \Gamma(A^*, \eta)$.

4.1 The SerVCS Approach

With the objective of enabling ontology development in distributed environments, where sets of changes are performed (cf. Definition 1) using different editors, the detection of *False-Positive Conflicts* (cf. Definition 4) by the VCS must be avoided. For this reason, ontologies should have a *unique serialization* (see cf. Definition 3). In order to realize that, we developed *SerVCS*, which generates a unique serialization of ontologies regardless of the used editing tool. The modeled concepts (triples) are ordered alphabetically in this unique serialization, first according to the *subject* name, then by *property* name. That way, ontologies (represented as text-based RDF documents) have always a consistent serialization in the remote repository. As a result, a high accuracy of conflict detection can be achieved and the identified conflicts are reduced to those caused by overlapping changes, *Syntactical Conflicts* (cf. Definition 2). This enables a VCS to automatically resolve most conflicts using its built-in algorithms. In the worst case, a user is confronted with conflicting changes and has to manually resolve them by providing a valid and consistent ontology. Since all ontologies have a unified serialization in the remote repository, the user is able to see the differences between any two versions of the ontology. Figure 2 illustrates the *SerVCS* approach, which consists of five main steps: (1) input: RDF documents serialized by different sorting criteria; (2) generate unique serialization; (3) output: RDF documents sorted with same criteria; (4) synchronization process from the point of view of VCS; and (5) final outcome: synchronized RDF document.

Figure 3 depicts the ontology development workflow using the *SerVCS* approach. After personal working copies are synchronized with the remote repository (cf. *Scene 1* of Fig. 1), users start performing their tasks using different ontology editors. When making any changes, such as adding, removing, or modifying existing concepts, the updated ontology is saved locally on the machine of the user, as illustrated in Fig. 2, *Scene 2* (which is still identical to *Scene 2* of Fig. 1). Next, these changes are uploaded to the remote repository. *Scene 3* shows that a unique serialization of the ontology is created as intermediate step. As a result, the concepts are organized using a common ordering criteria. In *Scene 4*, *User 1* uploads her changes successfully to the remote repository. Lastly, as illustrated in *Scene 5*, *User 2* starts uploading his changes to the remote repository. Since

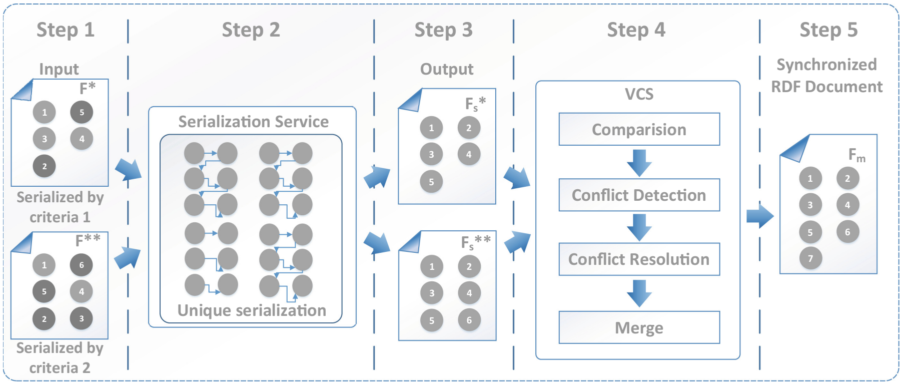


Fig. 2. The SerVCS approach. SerVCS receives RDF documents serialized by different sorting criteria (Step 1), and generates a synchronized RDF document (Step 5). In Step 2, a unique serialization is produced. RDF documents are sorted with same criteria (Step 3). Finally, a VCS synchronization process is performed (Step 4), i.e., comparison, conflict detection, conflict resolution, and merge.

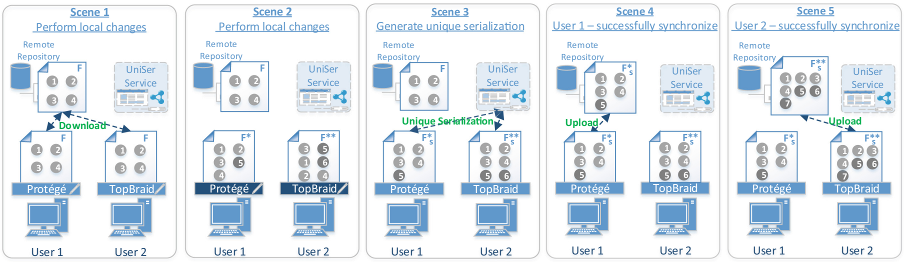


Fig. 3. The SerVCS development process. A distributed environment illustrating an ontology development process using SerVCS. Different ontology editors, e.g., Editors X and Y, are used for defining ontology F by Users 1 and 2. F* and F** represent local versions of F. Before synchronization with remote version, a unique serialization is created for F* and F**. F* is uploaded first. Next, changes in F** are successfully synchronized with F* since they have a unique serialization and any possible conflict is easy to be detected and resolved.

the ontology has a unified serialization, the VCS can merge both versions. In case of overlapping changes, the VCS shows exactly the lines which resulted in conflicts. Formally, a list of conflicts LC identified by *SerVCS* is defined as follows:

Definition 5 (List of Conflicts). *Given two RDF documents A and A^* such that F_1 and F_2 are serializations of A and A^* according to ordering criteria η_1 and η_2, a list $LC = <c_1, \ldots, c_n>$ of conflicts between F_1 and F_2, identified by* SerVCS, *comprises triples $c_i = (i, entry_{i1}, entry_{i2})$:*

- $i \in [1, MIN(size(F_1), size(F_2))]$,
- $entry_{i1} = (s_{i1}, p_{i1}, o_{i1})$ *and* $entry_{i2} = (s_{i2}, p_{i2}, o_{i2})$ *are RDF triples at the position* i *in* F_1 *and* F_2, *respectively,*
- $entry_{i1}$ *and* $entry_{i2}$ *are different, i.e.,* $s_{i1} \neq s_{i2}$ *or* $p_{i1} \neq p_{i2}$ *or* $o_{i1} \neq o_{i2}$.

Theorem 1. *Given serializations* F_1 *and* F_2 *according to ordering criteria* η *of RDF documents* A *and* A^*, *respectively. Consider* $LC = <c_1, \ldots, c_n>$ *the list of conflicts between* F_1 *and* F_2 *identified by* SerVCS. *If there are only syntactical conflicts between* A *and* A^*[2], *then for all* $c_i = (i, entry_{i1}, entry_{i2}) \in LC$

- $entry_{i1} = (s, p, o_{i1})$ *and* $entry_{i2} = (s, p, o_{i2})$, *and*
- $o_{i1} \neq o_{i2}$.

Proof. We proceed with a proof by contradiction. Assume that there are only syntactical conflicts between A and A^*, and there is a conflict c_i in LC, such that $c_i = (i, (s_{i1}, p_{i1}, o_{i1}), (s_{i2}, p_{i2}, o_{i2}))$, and $s_{i1} \neq s_{i2}$ or $p_{i1} \neq p_{i2}$. Since F_1 and F_2 are serializations according to the same ordering criteria η, $entry_{i1} \in \delta^-(A^*/A)$ and $entry_{i2} \in \delta^+(A^*/A)$. However, the statement $s_{i1} \neq s_{i2}$ or $p_{i1} \neq p_{i2}$ contradicts the fact that only syntactical conflicts exist between A and A^*.

5 Implementation

The architecture depicted in Fig. 4 has been implemented to empower VCSs to prevent wrongly indicated conflicts. The architecture consists of three main components: (1) a VCS, which handles different ontology versions via changesets; (2) a UniSer component, which generates unique serializations for the RDF documents; and (3) a repository hosting platform, which stores the RDF documents and propagates the changes.

5.1 Version Control System (VCS)

Git https://git-scm.com is used as the Version Control System, i.e., Git is responsible for managing different versions of the ontologies. Furthermore, the Git *hook* mechanism is used to automatize the process of generating the unique serialization of the ontologies before they are pushed to the remote repository. Once the modification of the ontology is finished, it is added to the Git *stage* phase. The next step proceeds with committing the current state to the personal working copy. The initialization of the commit event triggers a hook named

[2] Given two RDF-documents A and A^*, and $ChangeSet(A^*/A) = (\delta^+(A^*/A), \delta^-(A^*/A), <)$, there are only syntactical conflicts between A and A^*, iff $size(A^*) = size(A)$, and for each RDF triples t_i and t_j:

- $t_i \in \delta^-(A^*/A)$ and $t_j \in \delta^+(A^*/A)$,

then, there is a pair $(t_i, t_j) \in <$, and $t_i = (s, p, o_i)$, $t_j = (s, p, o_j)$, and $o_i \neq o_j$.

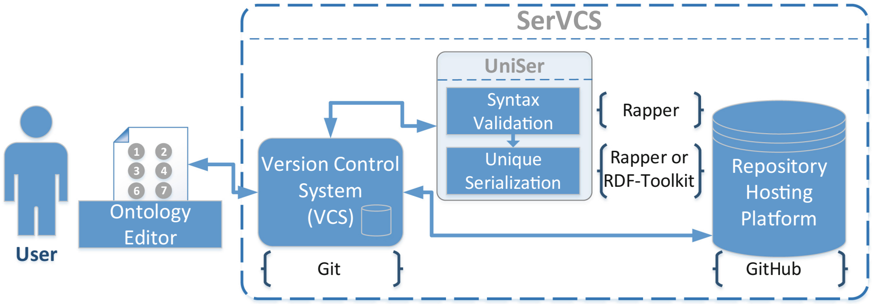


Fig. 4. The SerVCS architecture. Users interact with ontology editors, e.g., Protégé. (1) a VCS handles different ontology versions via changesets, e.g., Git. (2) The UniSer component performs syntax validation and generates unique serializations, e.g., Rapper or RDF-Toolkit. (3) A Repository Hosting Platform stores the ontologies and propagates the changes (GitHub).

pre-commit. This hook is adapted with a new workflow to handle the process of automatically generate a unique serialization, apart from the default one provided by Git.

SerVCS uses *Curl* https://curl.haxx.se as command-line HTTP client to send the modified files to the *UniSer* service. In case that ontologies fail to pass the integrated validation process, the commit is aborted and a corresponding error message is shown to the user. Otherwise, the files are organized according to the unique serialization. Subsequently, newly generated content overwrites the current content of the files by replacing the old serialization created by the ontology editor with the new unique serialization created by *UniSer*. When no error occurs during the entire process, the *pre-commit* hook event is completed and the commit is applied successfully. As a result, a new revision of the modified ontologies is created and the user is able to further proceed with successfully pushing her version to the remote repository. In addition, *Github* https://github.com is used as hosting platform for the repository to ease the collaborative development among several contributors.

5.2 UniSer

Furthermore, we implemented a stand-alone service, *UniSer*, using the cross-platform JavaScript runtime environment *Node.js* https://nodejs.org. Other tools are integrated to realize the tasks required for this service, e.g., syntax validation and unique serialization. The service accepts the ontology files as input through an HTTP interface and returns to the client either the error message from the validation process or the unique serialization of the file.

Once the input is received, *UniSer* validates the ontology, since a prerequisite for the unique serialization process is that ontology files are free of syntactic errors. The syntax validation is performed by *Rapper* http://librdf.org/raptor/rapper.html. In case of errors, a detailed report comprising the file name, error type, and error line is returned to the client. Otherwise, the process continues with creating a unique serialization using RDF-toolkit https://github.com/edmcouncil/rdf-toolkit or *Rapper* according to user preference for the sorting criteria to be used. During this task, a unified serialization of the ontology file is created by (1) grouping elements into categories, such as classes, properties, and instances, and (2) ordering elements within the categories alphabetically. The unique serialization of the ontology is send back to the client as final outcome.

In the following, we give some serializations of a simple ontology in *Turtle* https://www.w3.org/TR/turtle/ format comprising three concepts: *Train* and *UrbanTrain* of type *owl:Class* and a *UrbanTrain01* instance of class *Train*.

```
@prefix     : <http://example.com/> .
@prefix owl : <http://www.w3.org/2002/07/owl#> .
@prefix rdf : <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

:Train         rdf:type   owl:Class ;
               rdfs:comment "Train Concept"^^xs:string ;
               rdfs:label "Train"^^xs:string ;
               rdfs:subClassOf :Vehicle .

:UrbanTrain    rdf:type   owl:Class ;
               rdfs:comment "Train Concept"@en ;
               rdfs:label "Train"@en ;
               rdfs:subClassOf :Train .

:UrbanTrain01 rdf:type :UrbanTrain ;
               rdfs:comment "UrbanTrain01 operates in zone B"@en;
               rdfs:label "UrbanTrain01"@en .
```

The below excerpt serialized using Protégé tool is shown as follows:

```
@prefix     : <http://example.com/> .
@prefix owl : <http://www.w3.org/2002/07/owl#> .
@prefix rdf : <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

############################################
#        Classes
############################################

###      http://example.com/Train

:Train         rdf:type   owl:Class ;
               rdfs:comment "Train Concept"^^xs:string ;
               rdfs:label "Train"^^xs:string ;
               rdfs:subClassOf :Vehicle .

###      http://example.com/UrbanTrain
```

```
:UrbanTrain   rdf:type   owl:Class ;
              rdfs:comment "Train Concept"@en ;
              rdfs:label "Train"@en ;
              rdfs:subClassOf :Train .

#############################################
#        Individuals
#############################################

###      http://example.com/UrbanTrain01

:UrbanTrain01 rdf:type :UrbanTrain ;
              rdfs:comment "UrbanTrain01 operates in zone B"@en;
              rdfs:label "UrbanTrain01"@en .
```

The same ontology serialized with the TopBraid Composer is as follows:

```
@prefix     : <http://example.com/> .
@prefix owl : <http://www.w3.org/2002/07/owl#> .
@prefix rdf : <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

:Train
     a   owl:Class ;
     rdfs:comment "Train Concept"^^xs:string ;
     rdfs:label "Train"^^xs:string ;
     rdfs:subClassOf :Vehicle
     .

:UrbanTrain
     a   owl:Class ;
     rdfs:comment "Train Concept"@en ;
     rdfs:label "Train"@en ;
     rdfs:subClassOf :Train
     .

:UrbanTrain01
     a :UrbanTrain ;
     rdfs:comment "UrbanTrain01 operates in zone B"@en ;
     rdfs:label "UrbanTrain01"@en
     .
```

Using the *UniSer* service, the excerpt of the ontology is generated according to a unique serialization. The following listing depicts the result after the serialization by UniSer (which is nearly identical to the TopBraid Composer serialization).

```
@prefix     : <http://example.com/> .
@prefix owl : <http://www.w3.org/2002/07/owl#> .
```

```
@prefix rdf : <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

:Train         rdf:type   owl:Class ;
               rdfs:comment "Train Concept"^^xs:string ;
               rdfs:label "Train"^^xs:string ;
               rdfs:subClassOf :Vehicle .

:UrbanTrain    rdf:type   owl:Class ;
               rdfs:comment "Train Concept"@en ;
               rdfs:label "Train"@en ;
               rdfs:subClassOf :Train .

:UrbanTrain01 rdf:type :UrbanTrain ;
               rdfs:comment "UrbanTrain01 operates in zone B"@en;
               rdfs:label "UrbanTrain01"@en .
```

6 Experimental Study

In this section, we present the results of an experimental study investigating the effectiveness of the *SerVCS* approach. The goal of the experiment is to analyze the impact of ontology size, type of ontology changes, and sorting criteria on the behavior of *SerVCS*. We assess the following research questions:

RQ(1) Does the size of the ontology have an impact on the behavior of *SerVCS*?
RQ(2) Is the effectiveness of *SerVCS* affected by the ontology sorting criteria?

The experimental configuration to evaluate these research questions is as follows:

Ontologies: We compare the behavior of *SerVCS* using three ontologies of different sizes. Table 1 describes these ontologies in terms of number of triples, different subjects, properties, and objects.

- **Synthetic Ontology:** Synthetically generated small-size ontology composed of 16 RDF triples with six different subjects, four properties, and ten objects.
- **DBpedia Ontology** http://wiki.dbpedia.org/services-resources/ontology/**:** Medium-size ontology composed of 30,793 RDF triples. This ontology is used to describe Wikipedia infoboxes in DBpedia.
- **Gene Ontology** http://www.geneontology.org/ **(GO):** Large-size ontology composed of 1,540,109 RDF triples. GO describes molecular activities and relationships among genes.

Ontology Change Generation. Number of ontology changes are randomly generated following a *Poisson distribution*, i.e., we simulate ontology changes performed by users assuming that these changes obey a *Poisson distribution*.

Table 1. Ontology description. Ontologies of different sizes, described in terms of number of triples, subjects, properties, and objects.

Ontology	# triples	# subjects	# properties	# objects
Synthetic ontology	16	6	4	10
DBpedia ontology	30,793	3,986	23	16,807
Gene ontology	1,540,109	266,919	49	473,227

Table 2. Ontology changes [5]. Basic changes performed during ontology development process.

ID	Change type	Description	Example
CH1	Addition	Adding new elements like classes and properties	Add a new class, e.g., the class *Train* with properties *rdfs:label* and *rdfs:comment*
CH2	Modification	Modifying existing elements	Modify a property value, e.g., *rdfs:label* of *UrbanTrain* class
CH3	Deletion	Deleting existing elements	Delete an instance, e.g., the *UrbanTrain01* instance if it exists

The parameter λ indicates the average number of ontology changes per time interval, i.e., $\lambda = 2$ simulates that in average two ontology changes are performed per hour. Figure 5 illustrates the number and types of ontology changes performed by two users during eight hours. To ensure that our evaluation represents as much as possible a real usage scenario, a list of basic changes typically performed in ontology development is utilized (cf. Table 2). These changes are randomly chosen following a *uniform distribution* with replacement. We consider the change type *Modification* to be a combination of *Deletion* and *Addition*.

Metrics. We report on the number of conflicting lines (**NCL**). It is computed as the number of conflicts indicated by Git during the merge process of two versions of the ontology after each hour, and corresponds to the cardinality of the list of conflicts LC (cf. Definition 5).

Gold Standard. We compute the gold standard by summing up the number of conflicting lines (**NCL**), which corresponds to the cardinality of overlapping changes made by users in a specific hour (cf. Definition 2).

Implementation. Experiments were run on a Linux Ubuntu 14.04 machine with a 4th Gen Intel Core i5-4300U CPU, 3 MB Cache, 2.90 GHz with 8 GB RAM 1333MHz DDR3. *SerVCS* is implemented using Node.js version 4.4.5. The *syntax*

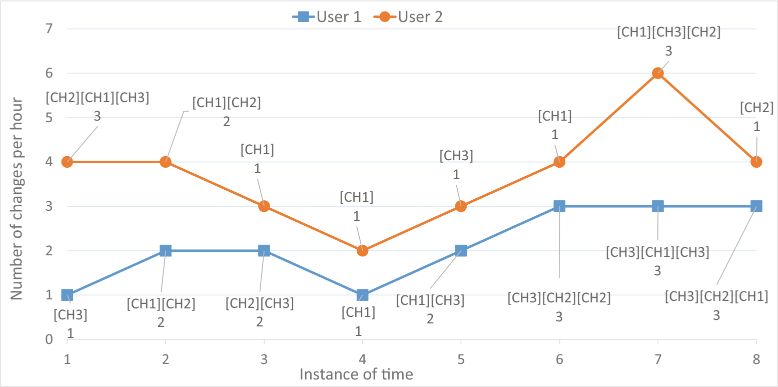


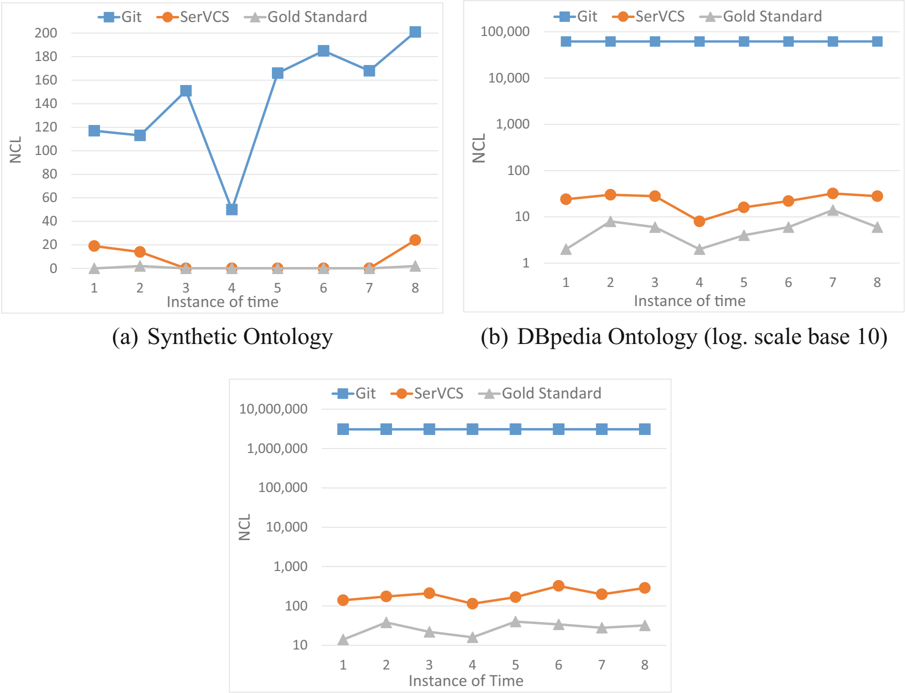
Fig. 5. Ontology change distribution [5]. Number and types of ontology changes (CH) per user in an interval of 8 hours. A Poisson distribution with $\lambda = 2$ models an average of two changes per hour. A uniform distribution with replacement is followed to sample the type of ontology changes (CH).

validation is realized using Rapper version 2.0.15 whereas the *unique serialization* is performed using RDF-toolkit version 1.4.0.1 and Rapper respectively. The used Git version is 1.9.1. The ontology change generator is implemented using RStudio version 0.99.902 https://www.rstudio.com/products/RStudio/.

Method. In order to answer the research questions, ontology changes of two users are simulated; two different ontology editors are assumed. *User 1* works with TopBraid, whereas *User 2* works with Protégé. Two scenarios are evaluated: (1) Users work purely with the functionalities of Git. (2) *SerVCS* along with Git as VCS is used. Log of ontology changes is kept during the experiment. In total, users make 30 changes: 11 additions, 9 modifications, and 10 deletions. The distribution of ontology changes per user is simulated with the Poisson distribution shown in Fig. 5. A log of ontology changes is available on GitHub https://github.com/lavdim/unistruct.

6.1 Impact of the Ontology Size

To answer research question **RQ1**, we follow the above described method to evaluate the behavior of plain Git and *SerVCS* with three different ontologies (cf. Table 1). Figure 6 shows the number of conflicting lines (NCL) in the *Gold Standard*, as well as the ones detected by Git and *SerVCS*. In the three ontologies, NCL values are significantly less in *SerVCS* than in Git. Moreover, in the Synthetic Ontology (small-size ontology), the NCL values are the same in *SerVCS* and *Gold Standard* for five time instances. In the DBpedia Ontology (medium-size ontology), *SerVCS* indicates up to three orders of magnitude less NCLs than Git. Finally, *SerVCS* reports up to four orders of magnitude less NCLs



(a) Synthetic Ontology (b) DBpedia Ontology (log. scale base 10)

(c) Gene Ontology (log. scale base 1,000)

Fig. 6. Impact of ontology size on SerVCS. Number of conflicting lines (**NCL**) detected by Git and *SerVCS* compared to the Gold Standard based on the Ontology Change Distribution in Fig. 5. (a) *SerVCS* detects the same NCLs as the Gold Standard in five instances of time in the Synthetic Ontology; (b) *SerVCS* indicates up to three orders of magnitude less NCLs than Git in the DBpedia Ontology; (c) *SerVCS* indicates up to four orders of magnitude less NCLs than Git in the Gene Ontology. *SerVCS* is not equally affected as Git.

than Git in the Gene Ontology (large-size ontology). Git utilizes a line-based algorithm for the comparison of ontology changes conducted by users. As the size of an ontology increases and modified ontologies are sorted differently, the number of compared ontology lines also increases. Therefore, Git performance is deteriorated as shown in Fig. 6. On the other hand, *SerVCS* compares pairs of changes ontologies also line-wise, but both documents are sorted using the same criteria and the space of potential conflicting lines in *SerVCS* is smaller. Thus, as shown in Fig. 6, *SerVCS* is not equally affected as Git from different ontology sizes. However, *SerVCS* may also wrongly identify conflicts, i.e., NCL values are not the same in *SerVCS* and *Gold Standard*. This behavior of *SerVCS* happens when users concurrently modify the subject or predicate of an RDF triple, i.e., a non-syntactical conflict is generated and Theorem 1 is not satisfied.

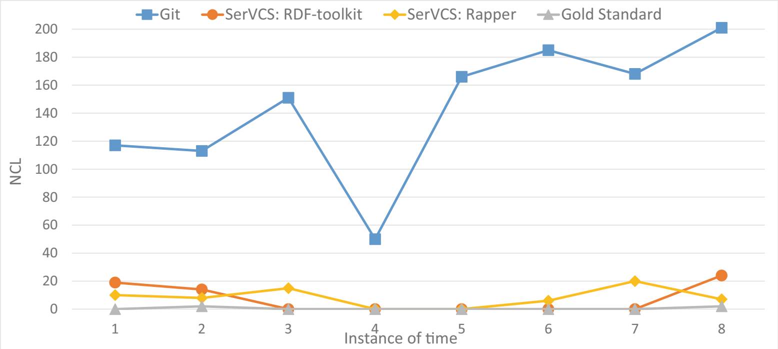


Fig. 7. Impact of sorting criteria. Number of conflicting lines (**NCL**) detected by Git and *SerVCS* compared to Gold Standard. Synthetic Ontology is modified according to the Ontology Change Distribution in Fig. 5. *SerVCS* follows two different sorting criteria produced by RDF-toolkit and Rapper. *SerVCS* exhibits similar behavior in both sorting criteria.

6.2 Impact of the Sorting Criteria

With the goal of answering research question **RQ2**, the experimental method is also followed when *SerVCS* utilizes different sorting criteria, i.e., RDF-toolkit and Rapper are used to generate unique serializations. RDF-toolkit sorts triples based on ontological concepts; RDF triples of classes, properties, and instances are categorized, and then, RDF triples are ordered alphabetically within category. Rapper simply sorts RDF triples in alphabetical order. Results in Fig. 7 show that *SerVCS* exhibits similar behavior in both sorting criteria and is able to identify the same NCL values as the *Gold Standard* in several instance times. Moreover, *SerVCS* also outperforms Git independently of the sorting criteria.

7 Conclusions and Future Work

This paper presents *SerVCS*, an approach for enabling VCSs to deal with various serializations of the same ontology in a multi-editor scenario. *SerVCS* relies on a unique serialization of ontologies with the objective of reducing the number of false-positive conflicts indicated by VCSs. Thus, users are enabled to develop ontologies in a distributed environment using different ontology editors. To study the effectiveness of *SerVCS* compared to Git, we performed an empirical evaluation. The results suggest that *SerVCS* reduces the number of false-positive conflicts when different ontology editors are utilized concurrently during the development process. Additional experiments were conducted to evaluate the impact of ontology size, as well as different sorting criteria. Effectiveness of *SerVCS* seems to be less impacted than Git whenever the size of the ontologies is increased or different sorting criteria are used to generate unique serializations.

As for future work, we envision to develop new techniques for automatically detecting and resolving not only syntactic but also semantic conflicts. We plan to empower *SerVCS* with probabilistic models on machine learning frameworks, and provide thus more accurate conflict detection and resolution strategies. Furthermore, a semantic layer will be added to *SerVCS* to prevent semantic inconsistencies that can be caused as a result of merging two versions of the same ontology. Finally, an extensive evaluation of the effectiveness and efficiency of the extended approach on both syntactic and semantic levels will be conducted.

Acknowledgements. This work has been supported by the European Union's Horizon 2020 programme for the project BigDataEurope (grant no. 644564), and the German Federal Ministry of Education and Research (BMBF) for the projects Industrial Data Space (grant no. 01IS15054) and SDI-X (grant no. 01IS15035C).

References

1. Palma, R., Corcho, Ó., Gómez-Pérez, A., Haase, P.: A holistic approach to collaborative ontology development based on change management. J. Web Semant. **9**, 299–314 (2011)
2. Halilaj, L., Grangel-González, I., Coskun, G., Lohmann, S., Auer, S.: Git4Voc: collaborative vocabulary development based on Git. Int. J. Semant. Comput. **10**, 167–192 (2016)
3. Mens, T.: A state-of-the-art survey on software merging. IEEE Trans. Softw. Eng. **28**, 449–462 (2002)
4. Altmanninger, K., Seidl, M., Wimmer, M.: A survey on model versioning approaches. Int. J. Web Inf. Syst. **5**, 271–304 (2009)
5. Halilaj, L., Grangel-González, I., Vidal, M., Lohmann, S., Auer, S.: Proactive prevention of false-positive conflicts in distributed ontology development. In: 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), vol. 2 (KEOD), pp. 43–51 (2016)
6. Dig, D., Manzoor, K., Johnson, R., Nguyen, T.N.: Refactoring-aware configuration management for object-oriented programs. In: 29th International Conference on Software Engineering (ICSE), pp. 427–436. IEEE (2007)
7. Ekman, T., Asklund, U.: Refactoring-aware versioning in eclipse. Electron. Notes Theor. Comput. Sci. **107**, 57–69 (2004)
8. Nguyen, H.V., Nguyen, M.H., Dang, S.C., Kästner, C., Nguyen, T.N.: Detecting semantic merge conflicts with variability-aware execution. In: 10th Joint Meeting on Foundations of Software Engineering (ESEC/FSE), pp. 926–929. ACM (2015)
9. Asenov, D., Guenat, B., Müller, P., Otth, M.: Precise version control of trees with line-based version control systems. In: Fundamental Approaches to Software Engineering (FASE) (2017, to appear)
10. Protzenko, J., Burckhardt, S., Moskal, M., McClurg, J.: Implementing real-time collaboration in TouchDevelop using AST merges. In: 3rd International Workshop on Mobile Development Lifecycle (MobileDeLi), pp. 25–27. ACM (2015)
11. Brun, Y., Holmes, R., Ernst, M.D., Notkin, D.: Proactive detection of collaboration conflicts. In: 19th ACM SIGSOFT Symposium on the Foundations of Software Engineering (FSE) and 13th European Software Engineering Conference (ESEC), pp. 168–178 (2011)
12. Lee, T.B., Connolly, D.: Delta: an ontology for the distribution of differences between RDF graphs. Technical report, W3C (2001)

13. Völkel, M., Groza, T.: SemVersion: an RDF-based ontology versioning system. In: IADIS International Conference on WWW/Internet (IADIS), IADIS, pp. 195–202 (2006)
14. Cassidy, S., Ballantine, J.: Version control for RDF triple stores. In: 2nd International Conference on Software and Data Technologies (ICSOFT). pp. 5–12 (2007)
15. Edwards, W.K.: Flexible conflict detection and management in collaborative applications. In: 10th Annual ACM Symposium on User Interface Software and Technology (UIST), pp. 139–148. ACM (1997)
16. Noy, N.F., Chugh, A., Liu, W., Musen, M.A.: A framework for ontology evolution in collaborative environments. In: 5th International Semantic Web Conference (ISWC), pp. 544–558 (2006)
17. Altmanninger, K.: Models in conflict - a semantically enhanced version control system for models. In: Doctoral Symposium at the ACM/IEEE 10th International Conference on Model-Driven Engineering Languages and Systems (MoDELS), CEUR-WS 262, CEUR-WS.org (2007)
18. Altmanninger, K., Schwinger, W., Kotsis, G.: Semantics for accurate conflict detection in SMoVer: specification, detection and presentation by example. In: Enterprise Information Systems and Advancing Business Solutions: Emerging Models, pp. 337–353. IGI Global (2012)
19. Brosch, P.: Improving conflict resolution in model versioning systems. In: Companion Volume of the 31st International Conference on Software Engineering (ICSE), pp. 355–358. IEEE (2009)
20. Cicchetti, A., Ruscio, D.D., Pierantonio, A.: Managing model conflicts in distributed development. In: 11th International Conference on Model Driven Engineering Languages and Systems (MoDELS), pp. 311–325 (2008)
21. Krusche, S., Brügge, B.: Model-based real-time synchronization. In: Softwaretechnik-Trends, vol. 34 (2014)
22. Chong, H., Zhang, R., Qin, Z.: Composite-based conflict resolution in merging versions of UML models. In: 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 127–132 (2016)
23. Gutierrez, C., Hurtado, C.A., Mendelzon, A.O., Pérez, J.: Foundations of semantic web databases. J. Comput. Syst. Sci. **77**, 520–541 (2011)

Conceptual Software Design: Modularity Matrix as Source of Conceptual Integrity

Iaakov Exman(✉)

The Jerusalem College of Engineering – Azrieli, Jerusalem, Israel
iaakov@jce.ac.il

Abstract. Conceptual Software Design is of utmost importance for software development due to its focus on the *Conceptual Integrity* of software systems. However, in order to turn it into actual standard practice in software design, a precise mathematical representation of Conceptual Design is necessary. This paper claims that Linear Software Models – by means of their basic algebraic structures, the Modularity Matrix or its corresponding Laplacian Matrix – guarantee Conceptual Integrity of the software system they represent. This is argued by first offering a concise Plausibility Path with a few formal steps towards *Conceptual Integrity* in terms of the Modularity Matrix. These steps clarify the role of the Modularity Matrix, both as a facilitator and as a formal source of the software modules' Conceptual Integrity. Then, the paper characterizes Conceptual Integrity as an intensive property of the software system. Finally, application in practice is demonstrated by providing explicit formulas to compute Conceptual Integrity principles, viz. propriety and orthogonality.

Keywords: Conceptual Software Design · Conceptual integrity
Linear Software Models · Modularity Matrix · Laplacian matrix
Modularity Lattice · Abstract Domain Conceptualization · Propriety
Orthogonality

1 Introduction

The original idea of *conceptual integrity* for software systems development was introduced by Brooks in his book "The Mythical Man-Month" [3]. There he argued for the utmost importance of *conceptual integrity* for software system design. We essentially agree with Brooks' qualitative statement, and claim that its formalization in mathematical terms should completely transform its practical applicability, actually enabling its usage for software systems.

This paper, an updated extension of [15] based in recent work, explicitly reformulates *conceptual integrity* in terms of Linear Software Models – our mathematical theory of software composition. In other words, the basic algebraic structures of this theory, viz. the Modularity Matrix by Exman [11] or its corresponding Laplacian Matrix by Exman and Sakhnini [14], guarantee the Conceptual Integrity of the software system they represent, by means of an iterative procedure. It is shown that the standard form of the Modularity Matrix is both the facilitator and a formal source of the software

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 233–248, 2019.
https://doi.org/10.1007/978-3-319-99701-8_11

modules' conceptual integrity. In case one deviates from the standard form, the matrix highlights the system spots in need of redesign, within the iterative procedure.

This Introduction concisely overviews the ideas of software *conceptual integrity* as presented by Brooks, and reviews the main Modularity Matrix properties.

1.1 Overview of Software Conceptual Integrity

In a more recent book by Brooks "The Design of Design: Essays of a computer scientist" [4], *conceptual integrity* was verbally described by three principles in terms of system functions. These principles are as follows:

Propriety – a software system contains only essential functions;
Orthogonality – functions are mutually independent;
Generality – many usage ways for each function.

The obstacle to practical application of these principles is that there have been no known translations to precise mathematical formulas and effective algorithms. This work provides the desired precise formalization of conceptual integrity, gaining both a deeper comprehension of Brooks' ideas and a clear basis for concrete usage of the referred principles.

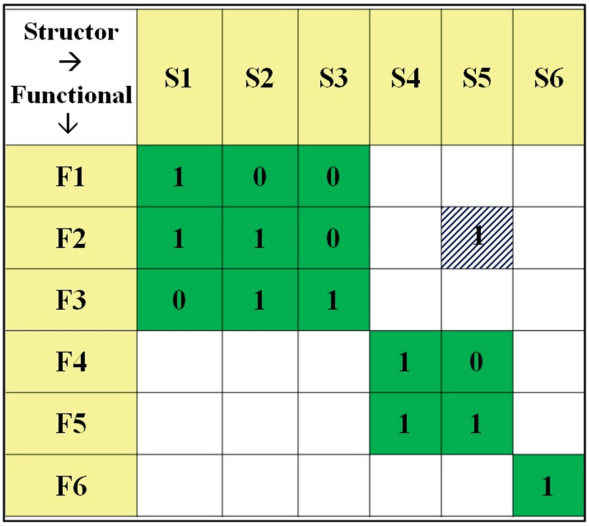
1.2 Modularity Matrix Concepts

The Modularity Matrix [9–11] enables to represent any level of a hierarchical software system, through sub-systems, to sub-sub-systems and so on, down to indivisible basic components. *Structors* – the matrix columns – stand for architectural structure units, generalizing classes of object-oriented languages. *Functionals* – the matrix rows – stand for architectural behavioral units, generalizing class functions, which may be invoked, but not necessarily.

Columns and/or rows reordering, together with algebraic manipulations [13], i.e. solving for the matrix eigenvectors, lead in the optimal situation to a square and block diagonal matrix. In this standard Modularity Matrix format, the blocks along the diagonal represent the *modules* of the current matrix level.

If there are outlier non-zero matrix elements beyond the boundaries of the diagonal modules, these outliers cause modules' coupling, which should be resolved by redesigning the system. This can be done, within an iterative procedure, by splitting modules or adding/removing structors and/or functionals. Figure 1 illustrates such an abstract Modularity Matrix, with one outlier matrix element, coupling two block-diagonal modules.

It has been shown by Exman and Sakhnini [14, 17], that a corresponding Laplacian Matrix can be generated from the Modularity Matrix. A similar procedure, also involving matrix eigenvectors, produces the same modules for the same system from both matrices. Thus, the Linear Software Models, indeed are a unified software composition theory.



Structor → Functional ↓	S1	S2	S3	S4	S5	S6
F1	1	0	0			
F2	1	1	0		1	
F3	0	1	1			
F4				1	0	
F5				1	1	
F6						1

Fig. 1. An abstract Modularity Matrix with an added outlier – A standard matrix means that it is strictly square and block-diagonal. This matrix is indeed square as it displays 6 structors (columns) and 6 functionals (rows). It is also almost block-diagonal, as it displays 3 modules seen as three blocks along the diagonal, (green background). A strictly block-diagonal matrix would have outside the modules (blank areas) only zero-valued matrix elements (values here omitted for increased clarity). But this matrix shows one outlier (in hatched dark blue background), a 1-valued matrix element in {F2, S5} coupling the top-left and the middle modules. This outlier hints at a need of software system redesign. (Color figure online)

1.3 Related Literature

Here we concisely review a sample of the related literature referring to Conceptual Integrity and algebraic structures, such as matrices and lattices which have been used for software system design.

Conceptual Integrity
Conceptual Integrity ideas for software design were first proposed in Frederick Brooks' books [3] and [4], as mentioned above in the beginning of this paper.

Jackson, starting from a research proposal [21], and co-authors elaborated Brooks' ideas by detailed explanations of case studies, e.g. on Git [7] and more recently De Rosso and Jackson [8]. Jackson has emphasized the importance of concepts for software systems, illustrating them by informal dependence graphs, from which simplified and more coherent subsets of concepts can be extracted [22].

Simplicity and regularity seem to be important characteristics of *Conceptual Integrity*. An example is a Technical Report by Kazman and Carriere [24], dealing with reconstruction of a software system architecture, using *conceptual integrity* as a guide. The architecture should in principle be built from small numbers of regularly connected components, with consistent functionality allocation to these components. Another example by Kazman [23] describes a SAAMtool, with visualization capability. *Conceptual Integrity* is estimated by the number of primitive patterns that a system uses.

Still another example is given by Clements et al. in their book [6], referring to conceptual integrity as a unifying design theme. The system should do similar things in similar ways, with small numbers of data and control mechanisms in the system. Issues with some similarity to our approach in this paper are: a - they mean the system at all hierarchical levels; b - a more precise definition of conceptual integrity would be given by counting mechanisms.

Occasional references concerning Conceptual Integrity have appeared in the literature. For instance, Beynon et al. [2] explicitly refer to Conceptual Integrity, but do not go beyond some vague statements about what it means. Orthogonality, one of the conceptual integrity principles, also have appeared in the software design literature. Krone and Snelting [25] refer to it in a paper using conceptual lattices extracted from source code.

Most recently, Exman and Katz [16] starting from an axiomatic approach, began to make explicit calculations with quantities expressing the Conceptual Integrity principles.

Algebraic Structures for Software System Design

Other algebraic structures, besides the Modularity Matrix, have been used for software systems design. The DSM (Design Structure Matrix) included in the Design Rules approach by Baldwin and Clark [1] has been applied mostly outside software engineering. It should be remarked that the DSM has been mostly analyzed by a superimposed economic options theory, external to the DSM itself, in contrast to our pure algebraic theory. For a set of references to this approach see e.g. [11].

Exman and Sakhnini [14, 17], have shown that a Laplacian matrix can be obtained from any Modularity Matrix, by means of an intermediate bipartite graph. Although a clearly different matrix, the Laplacian matrix obtains the same modules as its corresponding Modularity Matrix, by a similar spectral method – using eigenvectors and eigenvalues.

Another algebraic structure applicable to software system design is the Conceptual Lattice, developed within FCA (Formal Concept Analysis) mainly by Wille, Ganter and collaborators, see e.g. [19, 20]. It has been applied by a few authors to software analysis, see e.g. Krone and Snelting, [25]. More recently, Exman and Speicher [12] have shown the equivalence of the Modularity Lattice to the Modularity Matrix, displaying in alternative ways the same modules for any software system.

1.4 Paper Organization

The remaining of the paper is organized as follows. In Sect. 2 a Plausibility Path to conceptual integrity is offered. In Sect. 3 Conceptual Integrity is characterized as an intensive quantity of software. In Sect. 4 Conceptual Integrity is directly calculated from the Modularity Matrix. In Sect. 5 a discussion concludes the paper.

2 A Plausibility Path to Software Conceptual Integrity

We propose a Plausibility Path from an Abstract Domain Conceptualization to Software Conceptual Integrity. We assume that *Conceptual Integrity* pre-exists, in the abstract domain, before being formalized. We provide a general perspective of the plausibility path, leading through the Modularity Matrix to Software Conceptual Integrity. We then focus on each of its steps.

The main idea behind the Plausibility Path is to make plausible transitions between an acceptable starting point – the notion of abstract mathematical domain conceptualization – and the final goal of Software Conceptual Integrity. We call it Plausibility Path since we make acceptable statements in a heuristic fashion, but do not provide rigorous formal proofs.

We shall make formal definitions and corresponding calculation formulas in Sect. 4.

2.1 Plausibility Path Perspective

There are three essential formal steps from Abstract Conceptual Integrity to Software Conceptual Integrity, passing through the Modularity Matrix as shown in Fig. 2.

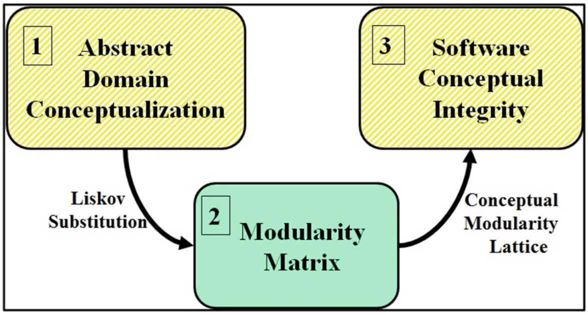

1 Abstract Domain Conceptualization
3 Software Conceptual Integrity
Liskov Substitution
2 Modularity Matrix
Conceptual Modularity Lattice

Fig. 2. From an Abstract Domain to Software Conceptual Integrity – The three steps are: the initial "Abstract Domain Conceptualization", the goal "Software Conceptual Integrity", and the intermediate Modularity Matrix. In between there are two transitions: "Liskov Substitution" and "Conceptual Modularity Lattice" to be later explained in the paper text.

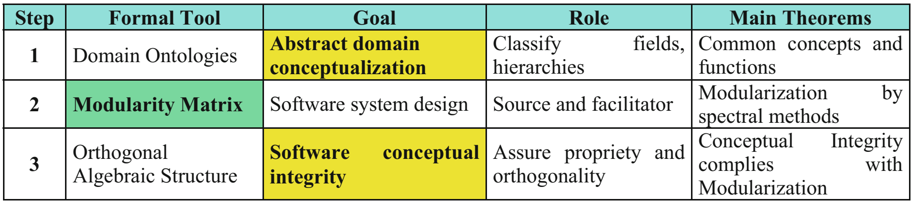
The meaning of the three formal steps is as follows:

1. *Abstract Domain Conceptualization* – along the history, concepts in e.g. Mathematics were grouped in fields within hierarchies obeying conceptual integrity;
2. *Modularity Matrix* – the basic algebraic structure of Linear Software Models, plays the role of both a facilitator and formal source for Conceptual Integrity;
3. *Software Conceptual Integrity* – the desired goal of the formalization steps, should assure software system orthogonality and propriety.

The meaning of the two transitions between the above steps is as follows:

1 → 2 – *Liskov Substitution* – translates abstract mathematical concepts into software entities;
2 → 3 – *Conceptual Modularity Lattice* – is an algebraic structure that has been shown to be equivalent to the Modularity Matrix, while obtaining concepts of the software modules.

One can summarize the roles of the above steps, as shown in Fig. 3.



Step	Formal Tool	Goal	Role	Main Theorems
1	Domain Ontologies	**Abstract domain conceptualization**	Classify fields, hierarchies	Common concepts and functions
2	**Modularity Matrix**	Software system design	Source and facilitator	Modularization by spectral methods
3	Orthogonal Algebraic Structure	**Software conceptual integrity**	Assure propriety and orthogonality	Conceptual Integrity complies with Modularization

Fig. 3. Plausibility Path: Tools, Goals and Roles – This summarizes properties of its formal steps in terms of their tools, goals and roles. See detailed discussion in subsequent subsections.

2.2 Software Structure and Behavior

Preliminary definitions clarify each of the above steps. The ultimate goal of the Plausibility Path is conceptual integrity in software systems. We refer to structure and behavior, thinking in terms of software, even when dealing with an abstract domain.

Definition A – ***Software Structure***
Software Structure is a relation among software architectural units ("structors", a generalization of classes) involving sub-classing and composition operators.

We use the same operators for software systems and for abstract ontologies. This follows common practice, emphasizing the analogies between abstract concepts and their respective software classes.

Definition B – ***Software Behavior***
Software Behavior is the performance of a function computation. The outcome of the function computation is a state change of the software system. We call "functionals" (a generalization of functions) the software architectural units of behavior.

Structors provide Functionals but the latter are not necessarily invoked. One often, by linguistic license, refers to the functionals themselves – without the performance of a computation – as software behavior.

2.3 Abstract Conceptual Integrity

Abstract concepts are hierarchically classified by properties' similarity. The hierarchy determines which concepts are particular cases of other ones. We illustrate the idea with some examples.

A square is a subclass of a rectangle, which is a subclass of a parallelogram. The parallelogram, the most general instance in this small hierarchy (in Fig. 4), is a polygon with four sides, in which the opposite sides are parallel. A rectangle is a subclass of a parallelogram with four right angles. A square is a subclass of a rectangle with all four sides equal.

Each lower hierarchy class has all the properties of the upper classes. A square has 4 sides (as in any quadrilateral), which are parallel (as in the parallelogram), and 4 right angles (as the rectangle).

Hierarchy is also true regarding behavior, i.e. the outcome of the functionals' calculations for each concept (or class). As an example, the perimeter of any class in this hierarchy is obtained by summing the length of the four sides (which in principle may be all different, partially different or all equal).

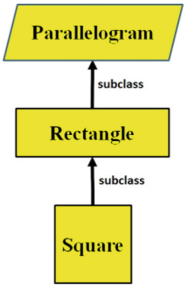


Fig. 4. The Quadrilateral Hierarchy – Each arrow (meaning subclass/subtype of) points from the particular concept (or class) to the more general concept. The parallelogram is the most general class of this hierarchy and Square is the most specific class. We visualize each class with the geometry of the class concept, instead of conventional UML rectangles for all classes.

A different hierarchy could contain a circle as a subclass of an ellipse. A yet different hierarchy would refer to 3-dimensional concepts such as a sphere as a subtype of an ellipsoid.

Each of the three referred hierarchies (quadrilaterals, ellipses, 3-D ellipsoids) display *conceptual integrity*, both intuitively and by some specific well-defined characteristic. For example, all quadrilaterals in Fig. 4 have linear segments as sides of a polygon (literally meaning "multiple angles"), while the ellipses have no linear segments and no angles in between at all the points in their perimeters.

These hierarchies, such as that the quadrilaterals in Fig. 4, are in fact small fragments of an ontology of geometric figures, see e.g. Rovetto [28], which may encompass the three referred hierarchies.

We summarize conceptual integrity in an abstract domain such as mathematics by means of the following Statement:

Statement 1 – ***Conceptual Integrity in Abstract Domain Hierarchy of Concepts***
In a class hierarchy determined by sub-classing, in an abstract domain, all the concepts of the hierarchy have at least one common concept, and one common function defined in the most general member of the hierarchy. The common concept and the common function represent the conceptual integrity.

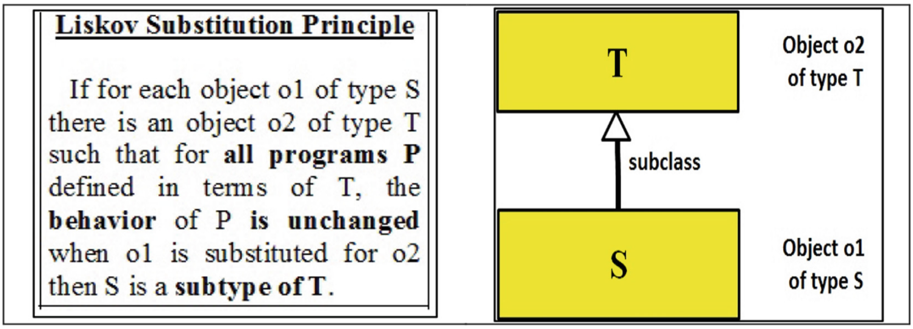
2.4 The Need for Liskov Substitution

We need Liskov Substitution to make the transition from an abstract domain, such as the quadrilaterals in Fig. 4, to actual software entities – say the same quadrilaterals which now have behavior, through their functionals, as represented by the Modularity Matrix. Thus Liskov Substitution attempts to translate, as faithfully as possible, concepts found in Abstract Mathematics to the software domain. This is possible, first of all, since the structure of both ontology fragments (hierarchies) in abstract mathematics and software hierarchies are based upon the same sub-typing operator.

The basic idea of Liskov Substitution which is relevant to *conceptual integrity* is to link "structure" to "behavior", effectively transforming concepts in an abstract (e.g. "mathematics") domain into generic software.

A formulation of Liskov Substitution [27] essentially links sub-classing to the behavior of software containing the relevant classes, when substituted by their subclasses. In such case, the software behavior should not change.

This is particularly interesting, as a "*structural*" class diagram (type T and its subtype S) in Fig. 5 is being linked to a "*behavioral*" condition, which is precisely what transforms an abstract domain to software concepts.



Liskov Substitution Principle

If for each object o1 of type S there is an object o2 of type T such that for **all programs P** defined in terms of T, the **behavior** of P **is unchanged** when o1 is substituted for o2 then S is a **subtype of T**.

Fig. 5. Liskov Substitution Principle and Class Diagram – The principle is shown in the left-hand-side of the figure. The class diagram illustrating Liskov's principle is in the right-hand-side of the figure. T is a class (or type). S is a subclass (or subtype) of T. Substitution of Object o2 of type T, by an object o1 of type S should not change the software behavior. This diagram is analogous to an abstract hierarchy in Fig. 4. Figure adapted from the paper by Exman [15].

2.5 The Modularity Matrix Roles

The Modularity Matrix of a software system is built of structors preserving the notion of sub-classing. Thus, the Modularity Matrix implicitly conveys the hierarchical ideas formulated in the previous sub-sections.

The contribution of the Modularity Matrix for Conceptual Integrity is to be purposely built to maximize modularity, increasing software simplicity and maximal orthogonality among modules.

Statement 2 – *Conceptual Integrity in the Modularity Matrix*
If the Modularity Matrix is standard (square and block-diagonal), then specific structors provide related functionals within modules, and the modules conceptual integrity is preserved for the restricted set of software systems represented by the Matrix.

2.6 Concepts in the Modularity Lattice

The contribution of the Conceptual Modularity Lattice in the transition from the Modularity Matrix to the software Conceptual Integrity is to link the optimization in terms of Structors and Functionals from the Modularity Matrix to concepts.

This is possible, since the Conceptual Modularity Lattice has been shown by Exman and Speicher [12] to convey information equivalent to the Modularity Matrix, in terms of software system modularity. Moreover, by its very definition from Formal Conceptual Analysis [19] the Conceptual Modularity Lattice is an algebraic structure restricted to the *concepts* relevant to its software system. This is summarized in the following statement.

Statement 3 – *Conceptual Integrity in the Modularity Lattice*
Since the Modularity Lattice in terms of software design is equivalent to its corresponding Modularity Matrix, the concepts fitting to the Matrix modules preserve conceptual integrity and this can be explicitly tested for the restricted set of software systems represented by the Modularity Lattice.

3 Conceptual Integrity as an Intensive Property of Software

In this section we go beyond the principles formulated upon the Modularity Matrix. We present and discuss the idea that Conceptual Integrity is an Intensive property of Software. We explain the meaning of intensive property, give an analogy to physical systems, and deal with software systems.

3.1 Conceptual Integrity Is an Intensive Quantity

Conceptual integrity, besides being a property of a whole hierarchical software system, seems to be a recursive property of each of its subsystems down to basic blocks. If any subsystem does not have conceptual integrity, it is plausible that the whole system cannot display it either.

We now give an example to explain what are intensive versus extensive quantities. Suppose that our system is a vehicle – either a car or a truck. A family car typically has 4 wheels. A truck usually has a bigger number of wheels. The weight of a vehicle is an *extensive* quantity: the weight of a vehicle is the sum of the weights of its parts. For instance, additional wheels increase the weight of the vehicle.

In contrast, the speed of a vehicle is an *intensive* quantity: the speed of the vehicle is not the sum of the speeds of its parts. All the car parts move at the same speed. Specifically, the tangential speed of any of its wheels is the same as the speed of the vehicle, irrespective of the number of wheels.

Conceptual Integrity is an intensive quantity. It is not the sum of the conceptual integrities of the components of a software system.

3.2 Increasing Conceptual Integrity by Exchange of Module Components

Here we use a different physical metaphor as a further illustration for the idea of Conceptual Integrity being intensive.

Assume a system having four sub-systems as in Fig. 6:

1. glass container;
2. water contained by the glass;
3. sphere mostly filled with air partially floating in the water;
4. small solid metal cube inside the sphere.

Now, one heats the glass container by an external heat bath. Heat energy flows among the different sub-systems, from those with higher temperatures to those with lower temperatures, until the whole system reaches a uniform temperature.

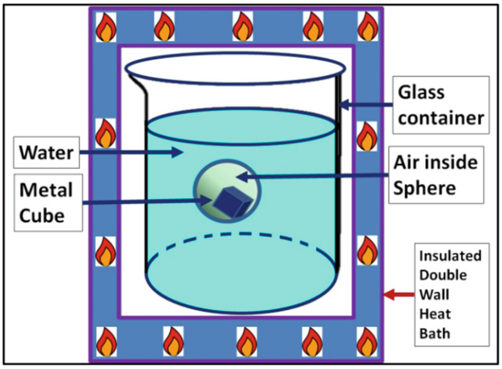


Fig. 6. Physical System Metaphor – The system has 4 sub-systems: a - glass container; b - water inside the glass; c - floating sphere filled with air; d - metal cube inside the sphere. A heat bath heats the glass container until the temperature is uniform, causing heat energy flow among the sub-systems. Figure reproduced from the paper by Exman [15].

In a software system, each sub-system usually has different computation characteristics – one dealing with data, another one with business logic, and so on. Moving concepts (classes) from one sub-system to another may increase conceptual integrity in both sub-systems. One could say that Conceptual Integrity in the whole system is optimized by flow of concepts (classes) among sub-systems. However, such flow of concepts and the intensive hypotheses of conceptual integrity, do not guarantee a single value of conceptual integrity anytime throughout a whole software system.

The previous physical metaphor suggests that *conceptual integrity* is not an extensive property, like heat energy, but an intensive property, like temperature.

4 Direct Computation of Conceptual Integrity from the Modularity Matrix

In this section we first assign a formal definition to two of the quantities behind Conceptual Integrity, viz. Propriety and Orthogonality. The proposed definitions are based on the criteria used to generate an optimized Modularity Matrix by an iterative procedure. Then we provide formulas to directly compute Conceptual Integrity quantities from the generated Modularity Matrix.

4.1 Propriety Formally Defined

Propriety has been verbalized in Subsect. 1.1 as “a software system contains only essential functions”. Intuitively, this means that one is minimizing the number of functions. Formally, it is stated as in the following definition.

Definition 1: Module Propriety

Propriety of a module in the Modularity Matrix, in a certain hierarchical level of a given software system, means that all its structors are mutually linearly independent, and concomitantly all its functionals are mutually linearly independent.

Explaining the previous intuition, the demand of ***linear independence*** among vectors (structors among themselves or functionals among themselves), implies that there are no two identical vectors. Moreover, if a sub-set of vectors are linearly dependent, some of the vectors are superfluous and can be eliminated. The decision of which vectors to eliminate is left to the software engineer, with a good knowledge of the concepts of the software system under development.

Thus, propriety reflects the fact that the Modularity Matrix optimizes – in fact minimizes – the number of structors and their provided functionals, for all its modules. Therefore, the above definition for a module extends to the whole matrix, and the standard Modularity Matrix complies with *Propriety*.

4.2 Orthogonality Formally Defined

Orthogonality was intuitively verbalized in Subsect. 1.1 as “functions are mutually independent”. Based upon the Modularity Matrix, this definition has two associated meanings:

- *Linear independence* – among structors and among functionals;
- *Strict orthogonality* – among modules, which is a stronger requirement than linear independence, and is easily visually recognized in the diagonal blocks of the Modularity Matrix, for instance in Fig. 1, when one ignores the outlier.

But, linear independence was already guaranteed by the Propriety definition 1. Therefore, the exact meaning of orthogonality is strict ***orthogonality*** among a specified sub-set of modules formally stated in the following definition.

Definition 2: Orthogonality among Modules
Orthogonality of a module with respect to a specified sub-set of other modules in the Modularity Matrix, in a certain hierarchical level of a given software system, means that all its structors are orthogonal to all the structors of the other modules in the specified sub-set, and concomitantly all its functionals are orthogonal to all the functionals of the other modules in the same specified sub-set.

The usage of the same term for the principle and for the linear algebra operation is not coincidence. It probably was suggested to conceptual integrity authors by the algebraic notion.

One could now ask about the third principle – Generality. Please see the Discussion Subsect. 5.1 item “d” for considerations on “Generality”.

4.3 Direct Computation of Propriety from the Modularity Matrix

Since *Propriety* has been defined in terms of linear independent vectors within a module, it is calculated, according to linear algebra, by the rank ***r*** of the sub-matrix of the given module. Specifically, if ***s*** is the number of structors (columns) of the module sub-matrix, propriety is calculated by Eq. (1).

$$\boldsymbol{Propriety} = \boldsymbol{1} - ((\boldsymbol{s} - \boldsymbol{r})/\boldsymbol{s}) \tag{1}$$

Note that since module sub-matrices are square, one could use as well the number of functionals ***f*** (rows) instead of the number of structors. The module propriety quantity in this equation has a value between zero and the maximum propriety value of 1 obtained when **r** equals **s**. The need for this kind of normalization is to facilitate calculations of propriety for the whole system, given the values for all its modules, while preserving Conceptual Integrity as an intensive property.

4.4 Direct Computation of Orthogonality from the Modularity Matrix

Orthogonality has been defined for all vectors within a module, with respect to vectors in other modules in specified sub-set of modules in the Modularity Matrix. According to linear algebra, orthogonality of a pair of vectors $\boldsymbol{v_{M1}}$ and $\boldsymbol{v_{M2}}$, respectively belonging to modules $\boldsymbol{M_1}$ and $\boldsymbol{M_2}$, is calculated by the scalar product of the pair of vectors in Eq. (2).

$$\boldsymbol{Orthogonality} = \boldsymbol{1} - (\boldsymbol{v_{M1}} \bullet \boldsymbol{v_{M2}}) \tag{2}$$

Note that each of the vectors in this equation is normalized (se e.g. Weisstein [29]), i.e. all their elements are divided by the length of the respective vector. Thus, the calculated orthogonality for a pair of vectors has a value between zero and the maximal value of 1 obtained for zero scalar product. Again the need for normalization is to facilitate calculations of orthogonality for the whole system, given the values for all pairs of structors and all pairs of functionals for all modules, preserving Conceptual Integrity as an intensive property.

5 Discussion

Conceptual Integrity has been considered of fundamental importance for software system design, but has been only vaguely defined.

This paper's basic claim is that the Modularity Matrix is a facilitator and a formal source of Conceptual Integrity information. We have provided two lines of argumentation:

a. *Plausibility Path from Abstract Domains through the Modularity Matrix to Conceptual Integrity* – We started from the accepted conceptual integrity of abstract domains, made a transition to the Modularity Matrix fitting a set of software systems. Using the equivalence to the Modularity Conceptual Lattice, we returned to "conceptual" aspects, to finally reach Conceptual Integrity.

b. *Formal definitions and direct calculation* – The Modularity Matrix optimization procedure was the direct source of the defined quantities viz. propriety and orthogonality, in a formal way.

Two of the principles – propriety and orthogonality – have a neat definition derived from the standard Modularity Matrix properties.

Promising progress has been achieved in this work, but additional investigation, in particular calculation for extensive numbers of case studies is needed to further clarify issues detailed in the next sub-section.

5.1 Open and Controversial Issues

a. Conceptual Independence of Abstract Hierarchies

We have referred in Sect. 2.3 to two independent hierarchies, one of *polygons* and another one of *ellipses*, say a circle. However, they are not strictly independent. One may think of a circle as a regular polygon in the limit of an infinite number of sides, enabling a transition between two of the above hierarchies. One can easily estimate the value of **π** in the perimeter of a circle **2 * π * Radius** by taking the limit of the perimeter of a polygon inscribed in the circle, when the number of polygon sides goes to infinity.

b. Stability Along Time of Conceptual Hierarchies
The situation is more complex than the naïve static view of Fig. 4 would suggest. Concepts evolve – see e.g. Lakatos [26] – in his book on "Proofs and Refutations" discussing the empirical contribution to the concept evolution of regular polyhedrons (from Euler's initial five). Concepts also can be said to expand along time – see e.g. Buzaglo [5] – according to the terminology of his book "The Logic of Concept Expansion".

c. The Single Brilliant Architect of Major Systems?
Brooks has argued in favor of the idea that only a single brilliant architect, can impart conceptual integrity to a major building, say an architect of a cathedral, or similarly to a major engineering enterprise such as a very large software system. Gabriel [18] challenges Brooks' position.

In our opinion, Brooks' position is difficult to be rationally proven for real systems. But its main drawback is the dependence on the existence and the opportunistic presence of a single brilliant mind. One obviously prefers a systematic construction of formal tools, based upon clear conceptual integrity ideas.

d. The Generality Principle of Conceptual Integrity
Generality, has been described as the quality that "a single function should be usable in many ways" in the same system. This intuitive formulation seems vague enough, being an obstacle to a formal interpretation. We shall return to this issue elsewhere.

5.2 Future Work

Open issues for future work include:

- extensive calculations on actual software systems;
- explanation of difficulties encountered with heavily used software systems such as Git [7].

5.3 Main Contribution

The main contribution of this work is that Linear Software Models – by means of the formal algebraic tools of Modularity Matrix or the Laplacian Matrix – guarantee Conceptual Integrity of the software system they represent.

References

1. Baldwin, C.Y., Clark, K.B.: Design Rules, Volume I. The Power of Modularity. MIT Press, Cambridge (2000)
2. Beynon, W.M., Boyatt, R.C., Chan, Z.E., Intuition in software development revisited. In: Proceedings of the 20th Annual Psychology of Programming Interest Group Conference, UK. Lancaster University (2008)

3. Brooks, F.P.: The Mythical Man-Month – Essays in Software Engineering, Anniversary edn. Addison-Wesley, Boston (1995)
4. Brooks, F.P.: The Design of Design: Essays from a Computer Scientist. Addison-Wesley, Boston (2010)
5. Buzaglo, M.: The Logic of Concept Expansion. Cambridge University Press, Cambridge (2002)
6. Clements, P., Kazman, R., Klein, M.: Evaluating Software Architecture: Methods and Case Studies. Addison-Wesley, Boston (2001)
7. De Rosso, S.P., Jackson, D.: What's wrong with git? A conceptual design analysis. In: Proceedings of Onward! Conference, pp. 37–51. ACM (2013). http://dx.doi.org/10.1145/2509578.2509584
8. De Rosso, S.P., Jackson, D.: Purposes, concepts, misfits, and a redesign of git. In: Proceedings of OOPSLA 2016, Conference, pp. 292–310. ACM (2016). http://dx.doi.org/10.1145/2983990.2984018
9. Exman, I.: Linear software models. In: Proceedings of GTSE 1st SEMAT Workshop on a General Theory of Software Engineering. KTH Royal Institute of Technology, Stockholm, Sweden, (2012). http://semat.org/wp-content/uploads/2012/10/GTSE_2012_Proceedings.pdf
10. Exman, I.: Linear software models, video presentation of paper [9] (2012). http://www.youtube.com/watch?v=EJfzArH8-ls
11. Exman, I.: Linear software models: standard modularity highlights residual coupling. Int. J. Softw. Eng. Knowl. Eng. **24**, 183–210 (2014). https://doi.org/10.1142/S0218194014500089
12. Exman, I., Speicher, D.: Linear software models: equivalence of modularity matrix to its modularity lattice. In: Proceedings of 10th ICSOFT International Conference on Software Technology, ScitePress, Portugal, pp. 109–116 (2015). https://doi.org/10.5220/0005557701090116
13. Exman, I.: Linear software models: decoupled modules from modularity matrix eigenvectors. Int. J. Softw. Eng. Knowl. Eng. **25**(8), 1395–1426 (2015). https://doi.org/10.1142/S0218194015500308
14. Exman, I., Sakhnini, R.: Linear software models: modularity analysis by the Laplacian matrix. In: Proceedings of ICSOFT 2016 11th International Joint Conference on Software Technologies, vol. 2, pp. 100–108 (2016). https://doi.org/10.5220/0005985601000108
15. Exman, I.: The modularity matrix as a source of software conceptual integrity. In: Proc. SKY'2016 7th International Workshop on Software Knowledge, ScitePress, Portugal, pp. 27–35 (2016). https://doi.org/10.5220/0006098300270035
16. Exman, I., Katz, P.: Conceptual software design: algebraic axioms for conceptual integrity. In: Proc. SEKE 2017, 29th International Conference on Software Engineering and Knowledge Engineering, pp. 155–160 (2017). http://dx.doi.org/10.18293/SEKE2017-148
17. Exman, I., Sakhnini, R.: Linear software models: bipartite isomorphism between Laplacian eigenvectors and modularity matrix eigenvectors. Int. J. Softw. Eng. Knowl. Eng. **28**(7), 897–935 (2018). https://doi.org/10.1142/S0218194018400107
18. Gabriel, R.P.: Designed as designer. In: Essay Track, ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages and Applications, Montreal, Canada (2007). http://dreamsongs.com/DesignedAsDesigner.html
19. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin (1998). https://doi.org/10.1007/978-3-642-59830-2
20. Ganter, B., Stumme, G., Wille, R.: Formal Concept Analysis - Foundations and Applications. Springer, Berlin (2005). https://doi.org/10.1007/978-3-540-31881-1
21. Jackson, D.: Conceptual design of software: a research agenda. CSAIL Technical report, MIT-CSAIL-TR-2013–020 (2013). http://dspace.mit.edu/bitstream/handle/1721.1/79826/MIT-CSAIL-TR-2013-020.pdf?sequence=2

22. Jackson, D.: Towards a theory of conceptual design for software. In: Proceedings Onward! 2015 ACM International Symposium on New Ideas, New Paradigms and Reflections on Programming & Software, pp. 282–296 (2015). https://doi.org/10.1145/2814228.2814248
23. Kazman, R.: Tool support for architecture analysis and design. In: ISAW 96 Proceedings of 2nd International Software Architecture Workshop, pp. 94–97. ACM, New York (1996). https://doi.org/10.1145/243327.243618
24. Kazman, R., Carriere, S.J.: Playing detective: reconstructing software architecture from available evidence. Technical report CMU/SEI-97-TR-010, Software Engineering Institute, Carnegie Mellon University, Pittsburgh (1997)
25. Krone, M., Snelting, G.: On the inference of configuration structures from source code. In: Proceedings of ICSE-16 16th International Conference on Software Engineering (1994). https://doi.org/10.1109/icse.1994.296765
26. Lakatos, I.: Proofs and Refutations: The Logic of Mathematical Discovery. Cambridge University Press, Cambridge (1976)
27. Liskov, B.: Keynote address - data abstraction and hierarchy. ACM SIGPLAN Not. **23**(5), 17–34 (1988). https://doi.org/10.1145/62139.62141
28. Rovetto, R.: The shape of shapes: an ontological exploration. In: Proceedings of SHAPES 1.0 1st Interdisciplinary Workshop on Shapes, Karlsruhe (2011)
29. Weisstein, E.W.: “Normalized vector” from MathWorld–a wolfram web resource (2018). http://mathworld.wolfram.com/NormalizedVector.html

Extraction of Patterns Using NLP: US and European Patents Domain

Anabel Fraga(✉), Juan Llorens, Eugenio Parra, and Valentín Moreno

Computer Science Department, Carlos III of Madrid University,
Av. Universidad 30, Leganés, Madrid, Spain
{afraga, llorens, vmpelayo}@inf.uc3m.es

Abstract. The job of reviewing patents applications might be complicated because every day the quantity of it is greater and greater. Also, the amount of work dedicated to preparing a proper application might be complicated. The process needs several revisions from investors and examiners. This revision job might have costs for the inventor because they don't know the proper mode for writing the application in the formal mode used. As part of a solution, one approach to minimize the impact of this fact and increase the success of the reviewing process is aid the human reviewer and also inventors with a set of patterns. The patterns are created using Natural Language Processing techniques and that accelerate the review just looking at the massive set of registration any similar application already patented. On the other hand aid the inventor in the process of writing an application in a formal manner.

Keywords: Indexing · Ontologies · Knowledge · Patterns · Reuse Retrieval · Patents · US patents · European patents

1 Introduction

The process for applying to an Intellectual Property protection, as patents, might be complex and reviewing your invention is really patentable must be approved and check by an examiner. Also, the language used to specify the invention in the application is specific to this domain.

If it could be possible to extract a set of patterns aiding the inventor and examiner in the process of construction of the application and also reviewing if the inventions could be already patented, the process of patenting could be improved in two different viewpoints [26, 27].

Christopher Manning states in his book that: "People write and say lots of different things, but the way people say things - even in drunken casual conversation - has some structure and regularity." [19]

In Requirements Engineering [16], Jeremy Dick introduced a type of pattern called boilerplates defined as "a language to express requirements." The author explained that a set of Boilerplates allow to collect and classify the different ways of expressing certain kinds of requirements.

Boilerplates, within the field of requirements engineering, have been defined as "a textual specification template of requirements, which are based on predefined patterns

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 249–264, 2019.
https://doi.org/10.1007/978-3-319-99701-8_12

in order to reduce ambiguity and ensure consistency in the way of expressing requirements" [9].

The important aspect in here is to ask ourselves: how do people write? Nowadays, researchers conduct investigations using natural language processing tools, generating indexing and semantic patterns that help to understand the structure and relation of how writers communicate through their papers.

This project will use a natural language processing system which will analyze a corpus of patents acquired from the open repository of the US patent and European patent Agencies. The documents will be processed by the system and will generate simple and composed patterns. These patterns will give us different results which we can analyze and conclude the common aspects the documents have even though they are created by different authors but are related to the same topic [2]. The study uses as the center of the study an ontology created in a nation founded project for Oncology and it has been extended with general terms of public health.

The remainder of this paper is as follows: Sect. 2 includes the state of the art and related work of the main topics of research, Sect. 3 includes the summary of the methodology; Sect. 4 summarizes the results, and finally conclusions.

2 State of the Art and Related Work

2.1 Information Reuse

Reuse in software engineering is present throughout the project life cycle, from the conceptual level to the definition and coding requirements. This concept is feasible to improve the quality and optimization of the project development, but it has difficulties in standardization of components and combination of features. Also, the software engineering discipline is constantly changing and updating, which quickly turns obsolete the reusable components [7, 18].

At the stage of system requirements reuse is implemented in templates to manage knowledge in a higher level of abstraction, providing advantages over lower levels and improving the quality of the project development. The patterns are fundamental reuse components that identify common characteristics between elements of a domain and can be incorporated into models or defined structures that can represent the knowledge in a better way.

2.2 Natural Language Processing

The need for implementing Natural Language Processing techniques (see Fig. 1) arises in the field of the human-machine interaction through many cases such as text mining, information extraction, language recognition, language translation, and text generation, fields that requires a lexical, syntactic and semantic analysis to be recognized by a computer [7]. The natural language processing consists of several stages which take into account the different techniques of analysis and classification supported by the current computer systems [8].

(1) Tokenization: The tokenization corresponds to a previous step on the analysis of the natural language processing, and its objective is to demarcate words by their sequences of characters grouped by their dependencies, using separators such as spaces and punctuation [6, 23]. Tokens are items that are standardized to improve their analysis and to simplify ambiguities in vocabulary and verbal tenses.

(2) Lexical Analysis: Lexical analysis aims to obtain standard tags for each word or token through a study that identifies the turning of vocabulary, such as gender, number and verbal irregularities of the candidate words. An efficient way to perform this analysis is by using a finite automaton that takes a repository of terms, relationships and equivalences between terms to make a conversion of a token to a standard format [15]. There are several additional approaches that use decision trees and unification of the databases for the lexical analysis but this not covered for this project implementation [31].

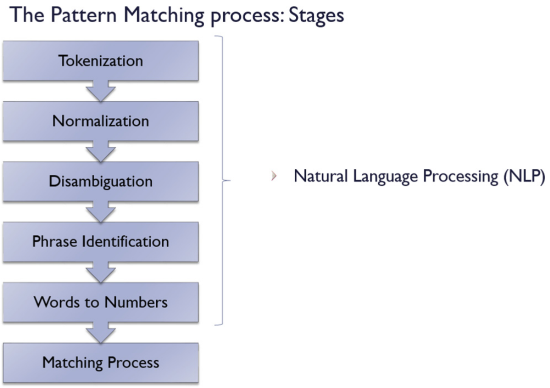


Fig. 1. NLP technique methods.

(3) Syntactic Analysis: The goal of the syntactic analysis is to explain the syntactic relations of texts to help a subsequent semantic interpretation [20], and thus using the relationships between terms in a proper context for an adequate normalization and standardization of terms. To incorporate lexical and syntactic analysis, in this project were used deductive techniques of standardization of terms that convert texts from a context defined by sentences through a special function or finite automata.

(4) Grammatical Tagging: Tagging is the process of assigning grammatical categories to terms of a text or corpus. Tags are defined into a dictionary of standard terms linked to grammatical categories (nouns, verbs, adverb, etc.), so it is important to normalize the terms before the tagging to avoid the use of non-standard terms. The most common issues of this process are about systems' poor performance (based on large corpus size), the identification of unknown terms for the dictionary, and ambiguities of words (same syntax but different meaning) [1, 32]. Grammatical tagging is a key factor in the identification and generation of semantic index patterns, in where the patterns consist of categories, not the terms themselves. The accuracy of this technique through the texts depends on the completeness and richness of the dictionary of grammatical tags.

(5) Semantic and Pragmatic Analysis: Semantic analysis aims to interpret the meaning of expressions, after on the results of the lexical and syntactic analysis. This analysis not only considers the semantics of the analyzed term but also considers the semantics of the contiguous terms within the same context. Automatic generation of index patterns at this stage and for this project does not consider the pragmatic analysis [25].

2.3 RSHP Model

RSHP is a model of information representation based on relationships that handle all types of artifacts (models, texts, codes, databases, etc.) using the same scheme. This model is used to store and link generated pattern lists to subsequently analyze them using specialized tools for knowledge representation [17]. Within the Knowledge Reuse Group at the University Carlos III of Madrid RSHP model is used for projects relevant to natural language processing [3, 12, 14, 29, 30]. The information model is presented in Fig. 2.

2.4 Ontology

The Ontology used in the research is an applied view of an Ontology, as shown in Fig. 3.

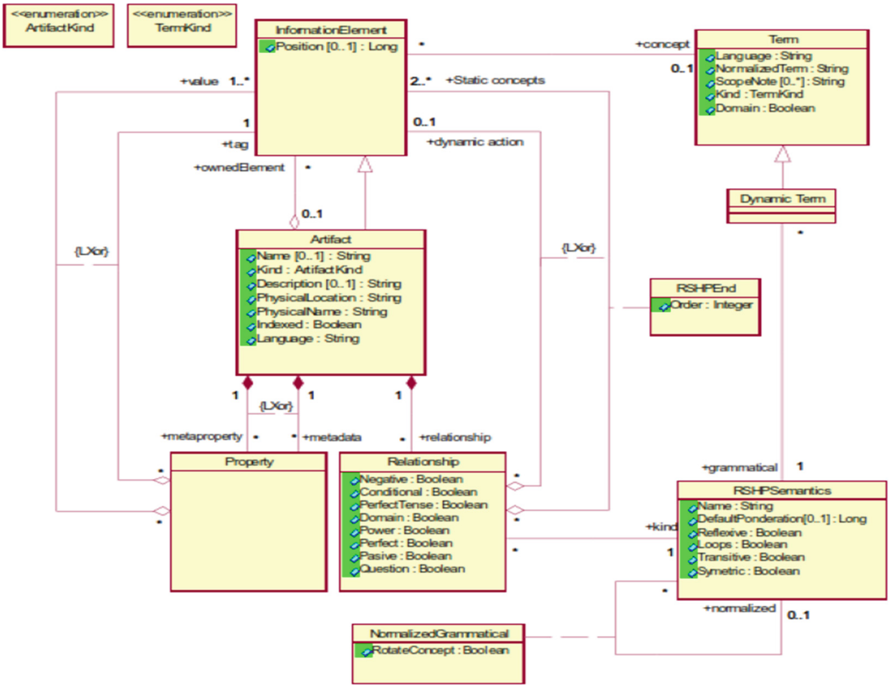


Fig. 2. RSHP information representation model [2, 13].

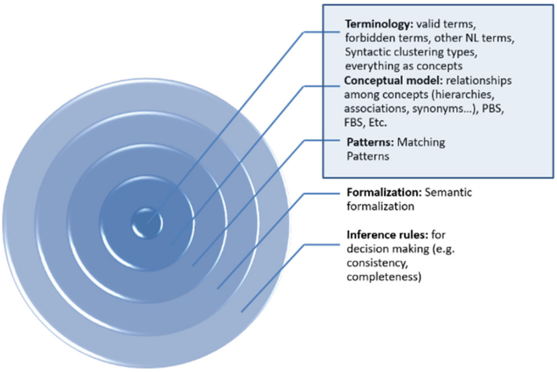

Terminology: valid terms, forbidden terms, other NL terms, Syntactic clustering types, everything as concepts
Conceptual model: relationships among concepts (hierarchies, associations, synonyms…), PBS, FBS, Etc.
Patterns: Matching Patterns
Formalization: Semantic formalization
Inference rules: for decision making (e.g. consistency, completeness)

Fig. 3. Ontology presentation and layers.

3 Methodology

The word Methodology is used frequently by researchers. Its meaning occasionally is not clear. People usually do not distinguish between methodology, process, method and technique; and moreover its meaning and differences. Let's make it clear in order to have an agreed terminology.

The word methodology is often erroneously considered synonymous with the word process [11, 21]. Martin's definitions for methodology, method, process, and tools are:

- "A Process (P) is a logical sequence of tasks performed to achieve a particular objective. A process defines what is to be done, without specifying how each task is performed. The structure of a process provides several levels of aggregation to allow analysis and definition to be done at various levels of detail to support different decision-making needs.
- A Method (M) consists of techniques for performing a task, in other words, it defines the how of each task. (In this context, the words method, technique, practice, and procedure are often used interchangeably.) At any level, process tasks are performed using methods. However, each method is also a process itself, with a sequence of tasks to be performed for that particular method. In other words, the how at one level of abstraction becomes the what at the next lower level.
- A Tool (T) is an instrument that, when applied to a particular method, can enhance the efficiency of the task; provided it is applied properly and by somebody with proper skills and training. The purpose of a tool should be to facilitate the accomplishment of the how. In a broader sense, a tool enhances the what and the how, most tools used to support systems engineering are computer- or software-based, which also known as Computer Aided Engineering (CAE) tools.
- Based on these definitions, a Methodology can be defined as a collection of related processes, methods, and tools. A methodology is essentially a recipe and can be thought of as the application of related processes, methods, and tools to a class of problems that all have something in common [5] ".

The objective of this research is to perform the extraction of syntactic-semantic patterns found within documents on patents.

Patent documents are written by experts, therefore we are saying that we will have very well written documents and high quality grammatical.

When the investigation is complete, we have a list sorted by frequency patterns (See Fig. 2). We will know the syntactic-semantic patterns that are most used when writing a patent.

In addition to patterns, the most recurrent words are known, we will identify the most common words in the patterns documents (Fig. 4).

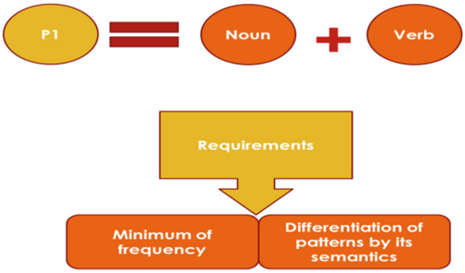


Fig. 4. The frequency of patterns [13].

Task 1: Representation of documents in syntactic and semantic categories.

The result of this task is the representation by syntactic and semantic categories of the complete content of the documents. The methods that compose the task are:

METHOD 1: Search for patent sources where they can download patents documents public and registered in PDF format. The documents must be converted to TXT format using pdf2txt. Pdf2txt is a program available on the internet.
METHOD 2: Download at least about 500 documents.
METHOD 3: Convert the PDF documents to TXT using the pdf2txt program.
METHOD 4: Get WordNet dictionary to form the ontology. This phase can be performed in parallel to steps 1, 2 and 3.

Task 2: Generating Basic Patterns

METHOD 5: Manage the ontology with a software for managing ontologies in the industrial domain, KnowledgeMANAGER. Adding vocabulary obtained in phase 4.
METHOD 6: Add the new ontology in BoilerPlates tool, a tool for detecting patterns (boilerplates in its most initial form) in a set of documents.

The BoilerPlates Tool is a software developed in a Ph.D. Dissertation [24] in order to generate patterns of text using Natural Language Processing solutions. The tool is an implementation of a patent [22] developed in the vb.Net language. In the first conception of the implementation, it was aimed at the automatic generation of patterns on specifications of engineering requirements. The process was developed taking into account all the syntactic and semantic categories contained in the ontology to which it was connected.

In order to improve the set of patterns resulting from the process, the functionality of the tool was expanded allowing the selection of the syntactic categories that were part of the process. In addition, the stop condition was parameterized, allowing the option of differentiating patterns by their semantics. All these measures allowed to experience the best conditions to obtain smaller sets of patterns but that could represent the information of the same requirements.

To allow the use of the pattern generation process in other domains, the tool was modified to obtain documents in plain text format as the input source.

In the last version of the tool, the substitution of patterns deleted by optional elements and wildcards, and an interface of the presentation of the results were added.

This tool has allowed doing part of the experimentation of the doctoral thesis [24], as well as end-of-course projects and final master's work [4, 10, 28].

METHOD 7: Define study scenarios and using ontology created, generating patterns with the BoilerPlates tool.
METHOD 8: TXT documents will be included one by one on the BoilerPlates tool, with this first step in the tool will generate the basic patterns.

Task 3: Pattern generation and analysis of results

METHOD 9: Representing one to one each scenario in BoilerPlates tool and start pattern generation.
METHOD 10: Analyze the results obtained by scenario.
METHOD 11: Analyze and compare the results of all scenarios.

In this work, a syntactic-semantic analysis is performed, of a sample of registered patents and made public, through an ontology based on natural language words (Fig. 5).

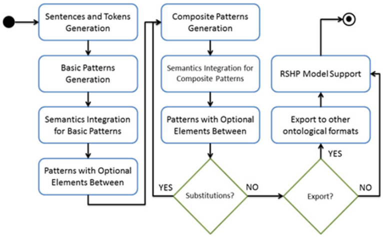

Sentences and Tokens Generation
Basic Patterns Generation
Semantics Integration for Basic Patterns
Patterns with Optional Elements Between
Composite Patterns Generation
Semantics Integration for Composite Patterns
Patterns with Optional Elements Between
YES
Substitutions?
NO
Export?
NO
YES
Export to other ontological formats
RSHP Model Support

Fig. 5. Analysis of sequences and patterns flow.

To get a larger sample of patent documents to analyze them, it has decided to use English as the language of analysis. Therefore all patents that are used in this investigation will be written in the English language.

The information processing summary of the scenarios is as follows:

Scenario 1 (a) and 2 (a) (USPTO and EPO):

- Use a minimum frequency of 1 to create patterns
- Differentiate patterns by their semantics is disabled and enabled

Scenario 3 and 4 (USPTO and EPO):

- Use a minimum frequency of 20 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 5 and 6 (USPTO and EPO):

- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 7 and 8 (USPTO and EPO):

- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is disabled

In all the cases:

- Use all grammatical categories.

All patents are searched in Internet and document must be PDF formats.

It does not establish any particular subject and not any particular area of investigation, the investigation developed here is valid for all subjects.

We have two samples of patents, on one hand, analyze documents of the United States Patent and Trademark Office, we have 359 documents, and secondly analyze documents of the European Patent Office, we have 379 documents Europeans different.

The study will be made with over 700 patent documents, all documents are analyzed with the BoilerPlates tool.

The ontology that includes the boilerplates tool, will be managed with the knowledge manager tool of REUSE Company. The vocabulary will form the ontology is providing by WordNet.

WordNet is used as a basis for the ontology of data recovery, we will have a language general controlled (not specialized by subject) and to language English. Into the WordNet, we obtain nouns, verbs, adjectives, and adverbs.

The investigation done here is interesting because we discover how the pattern of professional experts document their investigations, findings, and studies.

Here art to documentation is analyzed, so important it is to have an idea as important is knowing it registered.

The patterns that are obtained in the investigation may be useful in the future to guide the new professionals in the time of writing or searching for similar patents.

4 Results

The Scenarios followed in the experiments are:

Scenario 1:

- Sample USPTO (the United States Patent and Trademark Office) patents.
- All grammatical categories available are used
- Use a minimum frequency of 1 to create patterns
- Differentiate patterns by their semantics is disabled.

Scenario 2:

- Sample USPTO patents.
- Use all grammatical categories.
- Use a minimum frequency of 1 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 3:

- Sample USPTO patents.
- Use all grammatical categories.
- Use a minimum frequency of 20 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 4:

- Sample EPO (European Patent Office) patents.
- Use all grammatical categories.
- Use a minimum frequency of 20 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 5:

- Sample USPTO patents
- Use all grammatical categories.
- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 6:

- Sample EPO patents.
- Use all grammatical categories.
- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 7:

- Sample USPTO patents.

- Use all grammatical categories.
- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is disabled.

Scenario 8:

- Sample EPO patents
- Use all grammatical categories.
- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is disabled.

In addition to analyzing each scenario separately, comparisons between 1–2, 3–4, 5–6 and 7–8 pairs were carried out to compare the two sources of information used.

Comparative analysis - all in common scenarios will also be made to draw general conclusions to all the analyzed scenarios.

5 Basic Patterns

After the basic patterns were created, all the sentences from the text documents were analyzed and to each of the words (known in the database as token text), a term tag or syntactic tag was assigned with the help of the tables Rules Families and Vocabulary in the Requirements Classification database.

You may find the most repeated words in the domain of documents in the Basic patterns table. The most repeated words in grammatical categories such as nouns, verbs, and nouns coming from the ontology we used (Figs. 6, 7).

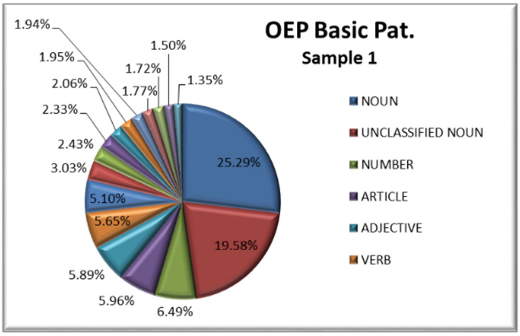


Fig. 6. Basic pattern results. OEP sample 1 [13].

It can be seen that in both cases there is little difference between the two samples, number of words in sample 1 than in sample 2, but not shown in the percentage of appearance in each of the grammatical categories (Fig. 8).

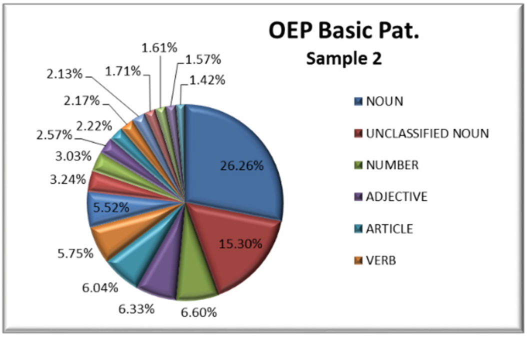


Fig. 7. Basic pattern results. OEP sample 2 [13].

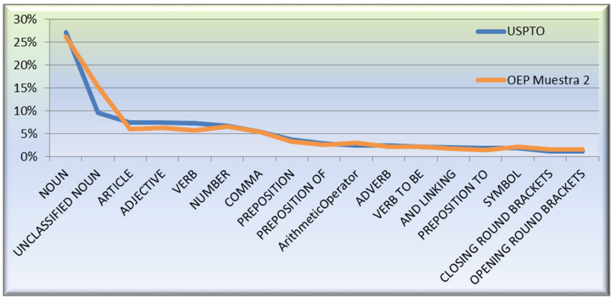


Fig. 8. USPTO vs OEP [13].

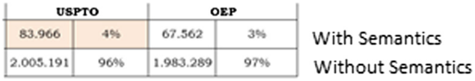
5.1 USPTO vs OEP

Comparing the two results, we see that the number of grammatical categories exceeding 1% is the same in both, but with slight differences. In USPTO items are the third most repeated grammatical category, while EPO are the numbers in this position. In the latter, the repetition of basic patterns is not rated much higher.

We see in the figure below the comparative representation of the 17 most repeated grammatical categories (Fig. 8).

5.2 Semantics

Semantics is present within the basic patterns but in a very limited way. We met a little more than semantics within American samples.



USPTO		OEP		
83.966	4%	67.562	3%	With Semantics
2.005.191	96%	1.983.289	97%	Without Semantics

In the next chart you can see the semantics that more appears in both samples: The stage 1, 2, 3, 5 and 7 are all made with American patent documents (Fig. 9).

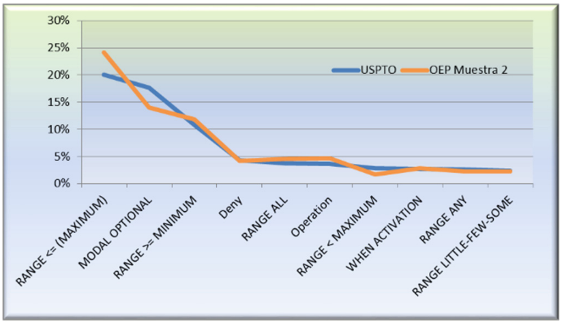


Fig. 9. Basic patterns: Semantics [13].

For the rest, we can conclude for the US shows the following is true:

- A higher minimum frequency, fewer patterns
- A higher minimum frequency, the lower the semantic obtained.
- Differentiate by semantic patterns is a better practice to know the real semantics being used when writing sentences. Otherwise, for the same pattern, which can adopt semantics could be anyone.
- Not differentiate by semantic results in increased number of patterns, but with fewer sub-patterns that form.

Scenarios 4, 6 and 8 are carried out with sample documents of European patents. We can conclude, for the European shows the following is true:

- A higher minimum frequency, fewer patterns
- A higher minimum frequency, the lower the semantic obtained.
- Differentiate by semantic patterns is a better practice to know the real semantics being used when writing sentences. Otherwise, for the same pattern, which can adopt semantics could be anyone.
- Not differ in semantics resulting in a greater number of patterns, and the number of subpatterns is very similar.

After the analysis of the US patents documents and European patents documents we can conclude the following:

The basic patterns obtained are independent of the frequency and the selection of grammatical categories in the boilerplates tool. All basic patterns are common within the same sample.

In the boilerplates tool, the higher the minimum frequency used, is less the number of patterns obtained and is shorter the time necessary to obtain them.

Differentiation has been made by their semantic patterns in the minimum frequencies of 1, 20 and 100 to US samples, and 20 and 100 for European samples. For frequency 1 it has not been possible to obtain results due to the high volume of information that we have handled. More than 25 days after running the tool, it has had to reject frequency 1 for the study. About the other two frequencies, we can say that the higher the frequency the number of patterns obtained is less.

Patterns are calculated without differentiation of semantics for the minimum frequencies of 1 and 100 with US sample. It is also calculated with the European sample for the minimum frequency of 100, without differentiation patterns by their semantics. It can be concluded that the same patterns are obtained with different semantics.

With increasing frequency, we lose patterns that have longer decomposition. Because the number of repetitions is less.

After using different frequencies to generate patterns in boilerplates, we can say that the intermediate frequency is what has given us the best results.

In both samples, the unclassified words are very present.

The patterns obtained in all scenarios can assist the writing for any user who needs to write a patent.

After the investigation, with the knowledge obtained now, we can give some recommendations to people who will do a similar study in the future.

The ontology can be improved, the ontology has 73 grammatical categories to define their vocabulary. For this project has not been completed because all the most important words are covered. The pending grammar to define is the type of punctuation, dates, email, arithmetic symbols, acronyms, etc. The undefined categories are shown in Fig. 8.

For future projects, scenarios of using a minimum frequency of 100 can be applied to search which is the minimum frequency that will create zero patterns.

It is possible to create a new analysis with a minimum frequency greater than 100 because we obtained patterns where their repetition frequency is greater than 100. But before begin studies with a higher minimum frequency, we recommend you should not consider words that do not correspond to a grammar of the ontology.

After ending all scenarios and analyzing results, we can conclude that authors writing papers about the same topic (in this case, genetic engineering) have similarity in how they write. They use a similar vocabulary and appropriate terms which makes the reading easier. Some additional enterprises need observation and intelligence.

6 Analysis of Results

Once the results has been analyzed, the documents of US patents and european patents, we can conclude the following:

- The basic patterns obtained are independent of the frequency and the selection of grammatical categories in the boilerplates tool. All basic patterns are common within the same sample.
- The higher the frequency used in the boilerplates tool, the smaller the number of patterns obtained and less time needed to obtain them.
- Differentiation has been made by their semantic patterns in the minimum frequencies of 1, 20 and 100 for American samples, and 20 and 100 for European samples. To frequency 1 it has not been possible to obtain results due to the large volume of information we have handled. After more than 25 days running the tool, it has had to dismiss frequency 1 for the study. On the other two frequencies, we can say that the higher the frequency the number of patterns obtained is lower.

- Patterns are calculated without differentiation of semantics for the minimum frequencies of 1 to 100 with American shows. It is also estimated with the European sample for the minimum rate of 100, without differentiation patterns by their semantics. It can be concluded that the same patterns are obtained with different semantics.
- By increasing the frequency lose patterns that have greater depth of decomposition. Since your number of repetitions is less.
- After using different frequencies to generate patterns in boilerplates, we can say that the intermediate frequency is what has given us better results.
- In both samples not rated names is very present.
- The patterns obtained in all scenarios may be of assistance to those who need to write a patent.
- After the investigation, with the knowledge now acquired, we can give some recommendations who faces a future in a similar study.
- The ontology can be improved, it has 73 labels for outstanding grammatical categories to define their vocabulary. For this project has not been completed because all the most important words are covered. The slopes are grammars to define the type of punctuation, dates, email, arithmetic symbols, acronyms, etc. These categories may be undefined in Fig. 8.
- There have been many token which are classified under the label "UNCLASSIFIED NOUN". For these cases we see three action plans:
- Or they could analyze them and give them all a grammatical category if possible, so finding patterns would be more accurate.
- If it is not possible to assign a particular category, you have to look at the possibility of eliminating all words and symbols are not classifiable.
- When generating patterns with boilerplates tool not consider the label "UNCLASSIFIED
- The documents have been used in this analysis can be improved by converting PDF to TXT performed in this process has been lost information. Documents with images are those that have lost more information.

It is possible to perform the analysis when the frequency is greater than 100 since we obtained patterns where the repetition frequency is greater than 100. But before studies with higher minimum frequency, it is recommended not to consider if the terms do not correspond to a grammar of the ontology.

7 Conclusion

"People write and say lots of different things,
but the way people say things -
even in drunken casual conversation -
has some structure and regularity".
-Christopher Manning

As Manning said and we are investigating in our hypothesis, it is possible to get structure and regularity in a set of documents in order to achieve a regular pattern to write or suggest better manners to prepare documents with a higher complexity. Natural language is complex and its regularity depends on the domain we are dealing with. The more regular and mature the domain, the best it is to extract a set of patterns. The more structure the more effective the set of patterns.

The structure of the Ontology is used in a generic composition, if adapted for each domain it could make more precise the set of patterns for each domain [13]. It is a future work to be performed in all the domains already analyzed: Health Care, Patents, and working right now in Banking.

Acknowledgment. The research leading to these results has received funding from the european union's seventh framework program (fp7/2007-2013) for crystal – critical system engineering acceleration joint undertaking under grant agreement n° 332830 and from specific national programs and/or funding authorities.

Also, the research leading to this paper has received funding from the AMASS project (H2020-ECSEL grant agreement no 692474; Spain's MINECO ref. PCIN-2015-262).

References

1. Abney, S.: Part-of-speech tagging and partial parsing. In: Young, S., Bloothooft, G. (eds.) Corpus-Based Methods in Language and Speech Processing. An Elsnet book. Bluwey Academic Publishers, Dordrecht (1997)
2. Alonso, L.: Herramientas libres para procesamiento del lenguaje natural. Facultad de matemática, astronomía y física. Unc, Córdoba, Argentina. 5tas Jornadas Regionales de Software Libre, 20 de Noviembre de 2005. http://www.cs.famaf.unc.edu.ar/~laura/freenlp
3. Amsler, R.A.: A taxonomy for english nouns and verbs. In: Proceedings of the 19th Annual Meeting of the Association for Computational Linguistic, Stanford, California, pp. 133–138 (1981)
4. Arroyo Minguela, L.: Extracción de patrones sintáctico-semánticos de documentos de patentes. Proyecto fin de carrera. Ingeniería técnica en informática de gestión. Escuela Politécnica Superior (2015)
5. Bloomberg, J., Schmelzer, R.: Service Orient or Be Doomed!: How Service Orientation will Change Your Business. Wiley, Hoboken (2006)
6. Carreras, X., Márquez, L.: Phrase recognition by filtering and ranking with perceptrons. In: Proceedings of the 4th RANLP Conference, Borovets, Bulgaria, September 2003
7. Cowie, J., Wilks, Y.: Information extraction. In: Dale, R. (ed.) Handbook of Natural Language Processing, pp. 241–260. Marcel Dekker, New York (2000)
8. Dale, R.: Symbolic approaches to natural language processing. In: Dale, R. (ed.) Handbook of Natural Language Processing. Marcel Dekker, New York (2000)
9. Daramola, O., Sindre, G., Stalhane, T.: Pattern-based security requirements specification using ontologies and boilerplates. In: 2012 Proceedings of 2nd IEEE International Workshop on Requirements Patterns, REPA 2012, pp. 54–59 (2012). https://doi.org/10.1109/repa.2012.6359973
10. De la o Maestro, N.: Evaluación de un sistema de procesamiento del lenguaje natural de la banca. Proyecto final de carrera. Ingeniería técnica en informática de gestión. Escuela Politécnica Superior (2015)
11. Estefan, J.A.: Survey of model-based systems engineering (MBSE) methodologies 2. Differentiating methodologies from processes, methods, and lifecycle models. Environment (2007). https://doi.org/10.1109/35.295942

12. Fraga, A.: A methodology for reusing any kind of knowledge: universal knowledge reuse. Ph.D. dissertation. Universidad Carlos III de Madrid (2010)
13. Fraga, A., et al.: Syntactic-semantic extraction of patterns applied to the us and european patents domain. SKY2016/ IC3K2016, Portugal (2016)
14. Gómez-Pérez, A., Fernando-López, M., Oscar, C.: Ontological Engineering. Springer, London (2004)
15. Hopcroft, J.E., Ullman, J.D.: Introduction to Automata Theory, Languages and Computations. Addison-Wesley, Reading (1979)
16. Hull, E., Jackson, K., Dick, J.: Requirements Engineering. Springer, Berlin (2010)
17. Llorens, J., Morato, J., Genova, G.: RSHP: an information representation model based on relationships. In: Damiani, E., Jain, L.C., Madravio, M. (eds.) Soft Computing in Software Engineering (Studies in Fuzziness and Soft Computing Series, vol. 159), pp. 221–253. Springer, Berlin (2004)
18. Llorens, J.: Definición de una metodología y una estructura de repositorio orientadas a la reutilización: el tesauro de software. Universidad Carlos III (1996)
19. Manning, C.: Foundations of Statistic Natural Language Processing, vol. 81. Cambridge University, Cambridge (1999)
20. Martí, M.A., Llisterri, J.: Tratamiento del lenguaje natural, p. 207. Universitat de Barcelona, Barcelona (2002)
21. Martin, J.N.: Systems Engineering Guidebook: A Process for Developing Systems and Products. CRC Press Inc., Boca Raton (1996)
22. Moreno, V., Suárez, P.M., Fraga, A., Llorens, J., Parra, E.: Método de generación de patrones semánticos. Pct/es2013/070638, Issued 2013 (2013)
23. Moreno, V.: Representación del conocimiento de proyectos de software mediante técnicas automatizadas. Anteproyecto de tesis doctoral. Universidad Carlos III de Madrid. Marzo (2009)
24. Parra, E.: Metodología orientada a la optimización automática de la calidad de los requisitos. Ph.D. (2016)
25. Poesio, M.: Semantic analysis. In: Dale, R. (ed.) Handbook of Natural Language Processing. Marcel Dekker, New York (2000)
26. Rehberg, C.P.: Automatic pattern generation in natural language processing. US Patent. US 8,180,629 b2, May 15, 2012, January 2010
27. Riley, M.D.: Some applications of tree-based modeling to speech and language indexing. In: Proceedings of the Darpa Speech and Natural Language Workshop, pp. 339–352. Morgan Kaufmann, California (1989)
28. Rodriguez Barberena, V.: Evaluation of a natural language processing system in public health. Trabajo final de máster. Máster en Ciencia y Tecnología Informática. Escuela Politécnica Superior
29. Suarez, P., Moreno, V., Fraga, A., Llorens, J.: Automatic generation of semantic patterns using techniques of natural language processing. In: SKY 2013, pp 34–44 (2013)
30. Thomason, R.H.: What is semantics? Version 2. http://web.eecs.umich.edu/~rthomaso/documents/general/what-is-semantics.html. Accessed 27 Mar 2012
31. Triviño, J.L., Morales Bueno, R.: A Spanish pos tagger with variable memory. In: Proceedings of the Sixth International Workshop on Parsing Technologies (IWPT-2000), Trento, Italia, pp. 254–265. ACL/SIGPARSE (2000)
32. Weischedel, R., Metter, M., Schwartz, R., Ramshaw, L., Palmucci, J.: Coping with ambiguity and unknown through probabilistic models. Comput Linguist **19**, 359–382 (1993)

Knowledge Management and Information Sharing

Cloud-Based Management of Machine Learning Generated Knowledge for Fleet Data Refinement

Petri Kannisto(✉) and David Hästbacka

Tampere University of Technology, P.O. Box 692, FI-33101 Tampere, Finland
{petri.kannisto,david.hastbacka}@tut.fi
http://www.tut.fi

Abstract. The modern mobile machinery has advanced on-board computer systems. They may execute various types of applications observing machine operation based on sensor data (such as feedback generators for more efficient operation). Measurement data utilisation requires preprocessing before use (e.g. outlier detection or dataset categorisation). As more and more data is collected from machine operation, better data preprocessing knowledge may be generated with data analyses. To enable the repeated deployment of that knowledge to machines in operation, information management must be considered; this is particularly challenging in geographically distributed fleets. This study considers both data refinement management and the refinement workflow required for data utilisation. The role of machine learning in data refinement knowledge generation is also considered. A functional cloud-managed data refinement component prototype has been implemented, and an experiment has been made with forestry data. The results indicate that the concept has considerable business potential.

Keywords: Distributed Knowledge Management · Mobile machinery
Cloud services · Data preprocessing · Machine learning

1 Introduction

The current era of industrial informatics has brought ever developing intelligent devices, data processing methods and sensor technology. Additional value can be gained from existing devices by collecting data and analysing it to have new information and knowledge. The importance of data analysis has been emphasised not only for business in general (LaValle et al. [19]) but also in industrial context (Duan and Xu [6]). For production, this also applies to mobile machines such as earthmoving, mining or forestry. Performance improvements not only bring competitive advantage but they also save resources and reduce emissions to the environment. In machinery, machine learning can be applied for multiple use cases. It may not only generate added value from data but it may also aid

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 267–286, 2019.
https://doi.org/10.1007/978-3-319-99701-8_13

the generation of data *preprocessing* knowledge that serves other data analysis tasks.

In this paper, a software concept is introduced – intended for service architectures – to enable the centralised management of fleet-wide sensor data refinement which is performed locally in mobile machines. The operation of modern machinery typically requires a high level of expertise, and even a skilled operator rarely has the technological knowledge required for optimal operation. That is, various feedback applications should be utilised to improve performance. In the data measured during operation, a lot of implicit information is available not only about the machine itself but also the material or the goods being processed. In an ecosystem, the number of machines and the amount of data can be arbitrarily large, and the machines may be geographically distributed. A centrally managed data refinement solution facilitates using all the potential of data as it unifies the information available for actual end user applications in various machines. The applications may, for instance, provide assistance in machine operation or adjustment. As data processing expertise and requirements are likely to evolve, frequent updates are expected.

This work has two main contributions: a conceptual cloud service architecture with machine learning and a functional prototype. The conceptual architecture utilises cloud services for storing extensive amounts of data as well as for machine learning to generate novel knowledge. The prototype covers an intermediary component that refines measurement data locally in machines – it receives its configuration from access points in a cloud so centralised management is achieved. The component accomplishes essential first-hand tasks thus generating information and facilitating further data analysis in end user applications.

The utilised research method is design science research. Novel knowledge is generated by designing artefacts that are evaluated against their requirements (referred to by e.g. March and Smith [22]).

This article is a revised version of a conference paper already published by the authors [15]. The original concept has been extended with cloud services and machine learning aspects, and some of the original contributions have also been more comprehensively explained.

The structure is as follows. Related work is discussed in Sect. 2. Section 3 discusses the actual problem followed by a solution design in Sect. 4. A forestry machine related prototype implementation is introduced in Sect. 5. Section 6 covers results and discussion while Sect. 7 concludes the paper.

2 Related Work

Among the publications in the industrial domain, no work has been found with a similarly extensive combination of a data processing workflow, cloud-based configurability and a data analysis or machine learning aspect. Various studies have been published with some common aspects though; thus, this part summarises the work related to either cloud services in production systems, machine data refinement, equipment data exchange or context awareness.

The Vehicular Cloud Computing (VCC) concept combines distributed data processing and mobility with a point of view different to this work. Its idea is to utilise the on-board computation and sensing capabilities of vehicles to enhance, for instance, traffic safety and management. Whaiduzzaman et al. [31] have written an extensive survey about the topic.

Storing machine or vehicle data in cloud is also a resource for large-scale data analysis. Bahga and Madisetti [1] have studied storing industrial measurement data in cloud to run analyses to raise maintenance performance. Filev et al. [8] show how vehicular data may be collected to a cloud and assisting services may be provided back to vehicles.

Even though both cloud and industrial production are related to this work, the Cloud Manufacturing concept is more related to business collaboration and interoperability within manufacturing networks. In manufacturing, the cloud service paradigm is expected to bring benefits such as scalability, agility and easier business networking. Tao et al. [28] have primarily envisioned manufacturing resource services while Wu et al. [32] have also covered product design as a cloud service.

Farming equipment related data collection or exchange has been researched in various papers. In a study by Steinberger et al. [26], farming equipment data is exposed in a service architecture. A work by Iftikhar and Pedersen [13] includes device data exchange in a bidirectional manner between office computers and farming machines. Peets et al. [25] provide a solution for data collection from various types of sensors. Fountas et al. [9] have introduced an information system concept for the management of farming machines. Machine data retrieval and integration concerns are present even in this work.

There are also other publications related to mobile machinery data processing. Palmroth [24] has studied the analysis of mobile machine data to assist operator learning. A knowledge management solution for operator performance assessment in the field is considered by Kannisto et al. [17]. Kannisto et al. [16] have introduced a system architecture to manage the information and knowledge required to assist machine parameter optimisation locally in machines. All of these studies contain machine data refinement, and the latter two have an information system architecture aspect. However, none of them has a similar level of detail in configurability, and cloud services have not been considered in the implementations.

Fault diagnostics and condition monitoring methods are related as they consider information generation by processing measured data. Various mathematical methods can be utilised for diagnostics as presented by Banerjee and Das, Basir and Yuan as well as Yang and Kim [2,3,33]. Condition-based maintenance (CBM) is enabled by utilising collected condition data as proposed by Jardine et al. [14]. Recently, even wireless sensor networks (WSN) have been utilised in diagnostics as suggested by Hou and Bergmann as well as Lu and Gungor [12,21]. These studies focus on data processing methods rather than knowledge management essential in this work.

Context recognition has been researched for a long time, and various methods as well as applications have been suggested. Khot et al. [18] provide a mathematical approach to recognising the context and the position of a tree planting robot; position information from various sources is combined mathematically to reduce error. Machinery is the domain also in the work of Golparvar-Fard et al. [10] where earthmoving equipment actions are recognised from video. Human activities recognition has also been researched including hospital work (Favela et al. [7]), car manufacturing (Stiefmeier et al. [27]) and general activities (Choudbury et al. [4]). Wan et al. [30] have even considered vehicular context recognition applications for parking assistance, vehicle routing and hazard prediction. In this paper, relatively little weight is put on context recognition so the method should not be compared with the advanced context recognition methods found in literature.

3 Data Processing Needs for Machine Fleets

3.1 Opportunities and Challenges of Data Analysis and Machine Learning

In the pursue for more efficient machine operation, this study recognises two data analysis use cases: fleet-wide and machine specific. The fleet-wide use case considers what is common for an entire group of machines. By utilising appropriate data analysis methods, multiple machine data sets may be processed together even if there were significant differences in machine types, operating environments and work types. In contrast, machine specific data analysis aims at discovering how a particular machine differs from the rest. Such differences appear due to the variation of machine parts: for instance, hydraulic components may vary even if they represented the same product, and a machine may have encountered more equipment wear than most similar machines.

To have a restricted scope, this work is concerned with the *information management* of fleet-wide data analysis for data preprocessing purposes. That is, while important, data analysis in individual machines, the actual data analysis methods as well as any end user applications are not in the scope (see Table 1). Still, end user applications bring the ultimate benefit to operators: the applications build added value by providing – for instance – assistance for machine operation.

Table 1. The scope of the work within machinery data utilisation.

Data analysis aspect	***Information management aspect***
Fleet-wide scope	Single machine scope
End user applications	***Data preprocessing***

While various data analysis methods could be utilised, this study emphasises the possibilities of machine learning. As huge amounts of operational machine

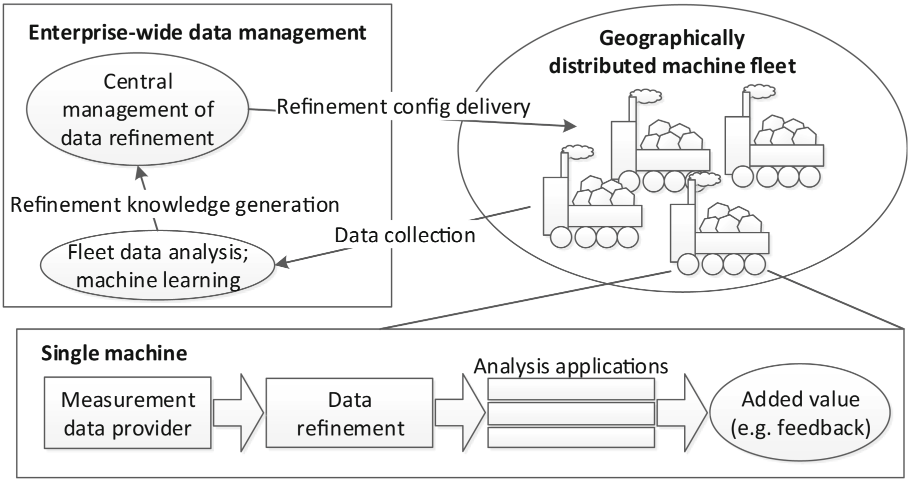

Enterprise-wide data management
Central management of data refinement
Refinement config delivery
Geographically distributed machine fleet
Refinement knowledge generation
Data collection
Fleet data analysis; machine learning
Single machine
Measurement data provider
Data refinement
Analysis applications
Added value (e.g. feedback)

Fig. 1. Data collection, analysis using fleet-wide data, refinement configuration management and data analysis results utilisation locally in machines. Adapted from [15].

fleet data are collected, machine learning methods may reveal new knowledge that might otherwise remain unobserved. That is due to the incomplete and ever-evolving nature of domain expertise – not only knowledge coverage improves but also new advances in machinery technology cause repeated changes. Especially, *deep learning* should be applied: it enables effective machine learning by utilising multiple abstraction levels (as stated by Deng and Yu [5, pp. 205–206] and LeCun et al. [20]).

Whichever are the end-user applications that utilise machine data, fleet-wide utilisation sets multiple requirements to data preprocessing and its management. Scalable configurability is essential: it must be possible to control data refinement even after it has been taken into use in a large geographically distributed fleet (see Fig. 1). Data collection enables analysis using fleet-wide data, producing data refinement configurations to be utilised locally in individual machines. The configurations are then utilised in data preprocessing for end user applications. Modern mobile machinery have advanced information systems and multiple sensors installed so a large amount of measurement data is produced during a work cycle, not to mention an entire work shift or weeks of operation. The more machines there are, the more data is generated. How distributed are the machines actually – may they operate anywhere in the world? What if the machines have no persistent internet connectivity?

This work is particularly motivated by forestry. Tree stem processing is a demanding task in terms of optimisation; there may be a lot of variation between forests and terrains even inside a geographically restricted area; also, there is often no internet connectivity in forests. Thus, the requirements of the next subsection give various forestry related examples.

3.2 Preprocessing Management Requirements

This study aims at data *preprocessing* management as it is typically required for sensor data. The scope is more extensive than just individual measurements as various more advanced features are beneficial for end user applications.

Data is structured as data sets called *data item collections*. A data item collection contains all the measurement values saved at a certain point in time. In addition, as modern machines have multiple operation related parameters (often customisable by the operator: such as the maximum power supplied to actuators), they are also stored. Thus, a data item collection provides a snapshot of machine state and performance. Data items are stored as a set of key-value pairs that enable access to data items using their identifiers. It is assumed that the machines of the same type have an identical key set in their data item collections. Once a data item collection has been retrieved, its items can be utilised for calculating or inferring derived data and information or to resolve the prevailing operating context.

In the forestry example, a data item collection represents the data of a single tree stem. For each stem, modern equipment supply various measurements such as felling diameter, stem length or how quickly the stem has been processed with the machine. The measurements and all machine parameters will be stored in a data item collection so stem data sets may be processed easily, one by one.

Machine type specific data item collection processing is likely required. First, variation is expected between machine types in measurement availability. For instance, as the degree of automation in tree stem processing keeps improving, a new machine model likely has more measurement items available compared to old ones. Second, models likely have variation in productivity, fuel consumption and other performance values. Third, variation in machine parametrisation is also expected due to differences in components such as hydraulic valves that control the machine boom and its implements. Parameter sets may vary as well as how a certain parameter affects machine operation.

Measurement failures must also be considered. Even a modern sensor may lack the ability to indicate if it has succeeded in measuring a value or not. Even if a sensor were not malfunctioning, there is still a possibility that its reading is not reliable – for instance, the sensor might have come off its installation position thus measuring something unexpected. In any case, it must be considered if each measurement value is reasonable or not. The motivation of outlier consideration has been discussed by, for instance, Osborne and Overbay [23].

Some variables cannot be measured as such but they have to be calculated. For instance, even if there were a measurement value for the productivity during a single work cycle, the daily productivity must be summed over a day. Further, a machine may change its position multiple times during a day, and working conditions may be so dirty that the windscreen must be cleaned multiple times during a work shift. If a productivity variable should only cover the actual material processing, any idle periods are to be excluded from productivity consideration. In forestry, an indirectly calculated Boolean flag may be utilised to inspect the tree species and size to help limiting data processing to a particular tree category.

As data item collections are persisted for later utilisation, each measurement value should be stored as such not to eliminate the possibility to recalculate values. This applies especially to cases where long-time historical data is required in analysis. If a measurement value is considered out of outlier limits and automatically declared a failure, it will be impossible to reprocess it in case of a later change in outlier conditions. Therefore, in many cases, it is a good practise *not* to store any values calculated from measurements as calculation formulas might evolve. Naturally, in some applications, if original values are not needed for sure, it may also be appropriate to save storage space by only saving the essential derived values rather than all raw values. However, if it is possible to submit data often to cloud, space is typically not a problem.

To run data analyses in a large scale, it is beneficial if data item collections are categorised. There may be considerable systematic variation in their values. Not to treat them as a homogeneous mass (what they certainly are not), at least rough categorisation is beneficial so each data item collection may be treated within an appropriate group. In forestry, each stem may be categorised after its size or tree species as it likely affects productivity – if the processing of large trees is being optimised, little trees should be ignored. As categorisation is performed based on measured values, it is subject to failures; it cannot be performed if some required value has been measured incorrectly.

Mobile machines may operate in varying environments so the power of context awareness should be exploited – the context may significantly affect how a machine can perform as argued by Väyrynen et al. [29]. Depending on the context, an absolute numeric value may be relatively high or low. It must be considered if performance value comparison is appropriate if the values have been measured in different contexts. For instance, performance is likely low in unfavourable conditions: the temperature may affect fuel consumption, rough terrain makes machine movement slower and so forth. In context classification, its subtleness and other aspects must be considered depending on the application area. Another important consideration is knowledge evolution: it may also be required to update the selected context classification method sometimes.

Context recognition is essential also in forestry. Even inside a relatively small geographical region, there may be a lot of variation between forests: the type of land may affect machine performance, and tree species may also vary. Also, the type of work being performed (final felling, thinning or other) always affects absolute productivity values.

Due to machine fleet distribution, data caching is important. First, the requirement applies to configuration delivery: the data refinement application cannot rely on network connectivity so it needs local copies for any configuration items. Second, as measurement data is collected from machines for future data analysis activity, similar caching is required so the data can wait for delivery to the enterprise cloud.

The various requirements and related specification items are summarised in Table 2. The required data preprocessing tasks cover e.g. data structures, indirect measure calculation and contextual variation.

Table 2. Data preprocessing requirements summarised.

Requirement	Conforming specification item
Associate related data in sets	Use data item collections
Machine types have differences	Consider machine types in data processing
Measurement errors reduce analysis reliability	Data outlier analysis
Indirectly calculated measures	Support for derived variables
Allow data calculation evolution for old data	Store raw measurement values as such
Distinction and grouping of data sets	Data item collection categorisation
Operating environment and work type differences	Context recognition support
No persistent internet connectivity	Data caching

4 Managing Data Refinement with Cloud Services and Machine Learning

4.1 Refinement Workflow

Considering given requirements, a solution can be designed. The flow of the application run locally in machines is illustrated in Fig. 2. There are four main phases complemented by context consideration. To enable the utilisation of constantly evolving domain expertise, some phases utilise externally defined methods or configuration files. Each phase is explained in the coming paragraphs.

First, measurement values are retrieved; they are stored in data item collections realised as key-value pairs. For a certain machine type, each collection is expected to have the same key-value pairs. In forestry, a reasonable data structure is to have a data item collection for each processed tree stem.

Then, an outlier check is performed. Whatever the utilised method is, it should be applied early as it may affect forthcoming data processing.

The next phase covers the calculation of derived variables (i.e. the data not directly measurable). Naturally, a derived variable cannot be calculated if any required measurement has failed. In this work, derived Boolean values associated to a data item collection are called *appearance* items: whether some condition set is fulfilled by the collection or not. For instance, in forestry data, an appearance item may express whether the species of a stem is spruce. The information may be utilised in further data analysis to easily determine which stems are interesting – for example, occasional birches in a spruce forest may be ignored.

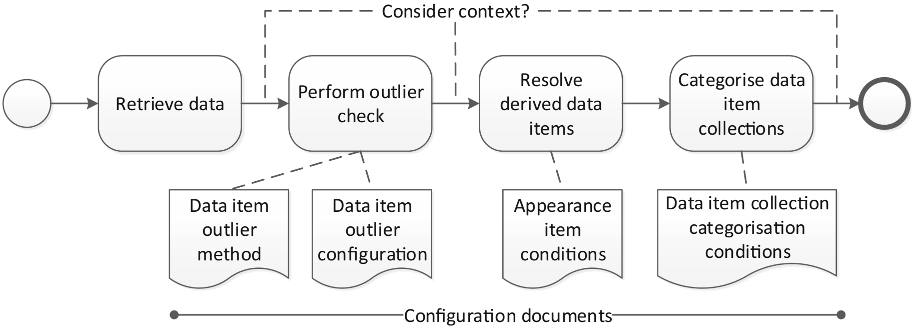

Consider context?
Retrieve data
Perform outlier check
Resolve derived data items
Categorise data item collections
Data item outlier method
Data item outlier configuration
Appearance item conditions
Data item collection categorisation conditions
Configuration documents

Fig. 2. Data refinement flow. Adapted from [15].

Finally, each data item collection is categorised. Whatever the categorisation criteria are, technically, they consist of condition sets on measurement values. If a data item collection has a failed measurement value that is required for categorisation, the collection is ignored in tasks where categories are essential.

Depending on the application, context awareness may be applied in several phases. Context information may even affect the outlier check; for instance, it may determine which numeric outlier limits are applied or it may determine what kind of outlier check method is utilised. Later in the refinement flow, the context may affect how derived data items are resolved. However, some context awareness methods may require data item collection categorisation results so they cannot be utilised earlier. In the end, even though the workflow has certain phases, its design is adaptable in terms of context awareness.

Let us consider forestry again to have a workflow execution example. First, an outlier check is required. For instance, if a measured value is beyond its reasonable limits, it must be declared a failure. Second, derived variables are calculated. Typical effectiveness variables (such as wood volume productivity while processing a single stem) are such as they cannot be measured directly. Also, some derived variables may require considering multiple data item collections (i.e. stems; such as the mass of processed wood per working hour during a day). Another derived variable could be the Boolean value (i.e. appearance item) whether a stem is "large" which involves the comparison of its felling diameter to a specific limit. Third, data item collections are categorised according to predefined conditions. Depending on the objective of the categorisation, stem categories could include tree species, tree sizes or both. Besides the mentioned phases, context-awareness may be applied in multiple parts in the flow. One option is simply to let the predominant tree stem category determine the prevailing context – this design choice depends on the application.

4.2 Cloud-Based Configuration Management with Machine Learning

Data refinement configuration management is illustrated in Fig. 3. The number of machines is arbitrary as well as their geographic locations. Various applications may utilise refined machine data, but the aspects of managing the actual refinement are explained in the following paragraphs.

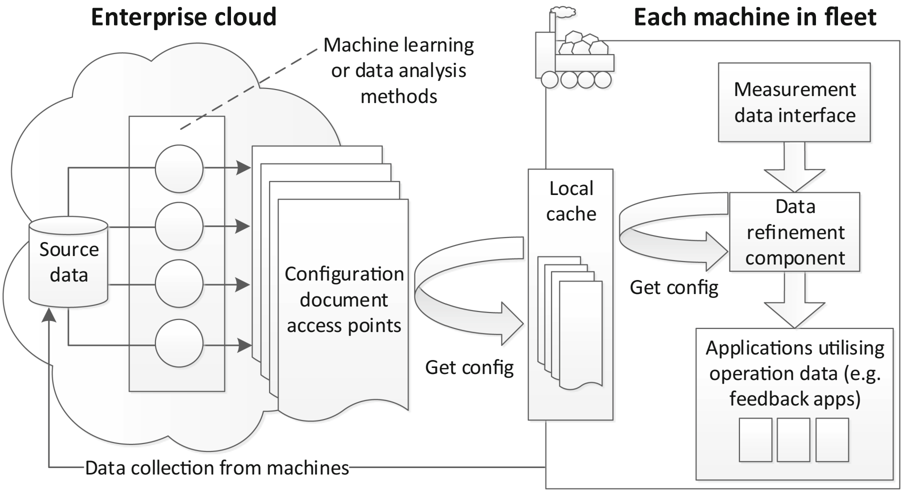

Enterprise cloud
Machine learning or data analysis methods
Source data
Configuration document access points
Get config
Each machine in fleet
Local cache
Get config
Measurement data interface
Data refinement component
Applications utilising operation data (e.g. feedback apps)
Data collection from machines

Fig. 3. Data analysis and data refinement management illustrated. Adapted from [15].

A software component has been designed to implement the data refinement workflow that utilises externally provided configuration documents. In each individual machine, it retrieves raw measurement data from the measurement data interface of the machine. Due to internet connection limitations, a cache holds local copies of the prevailing refinement configuration retrieved from the enterprise cloud. Having a software *component* enables reuse for the functionality in an arbitrary number of applications.

The enterprise cloud has multiple tasks in the data refinement management concept. First, it maintains a centralised storage for machine data. A large coverage is required for effective fleet-wide information generation. Second, utilising the stored data, machine learning or other data analysis methods are applied to generate the data refinement configuration utilised locally in machines. Multiple analysis methods are required as there are various configuration items. Third, the cloud stores the analysis results – i.e. the refinement configuration documents – and provides access points to make them available for machines. The everyday technology portfolio covers various networking methods for configuration retrieval such as HTTP (Hypertext Transfer Protocol) widely supported by software libraries. In the end, the cloud paradigm provides a basis for centralised management and scalable business in an environment where the data amount is huge and distribution requirements are ultimate.

5 Cloud-Enabled Data Refinement Prototype

5.1 Concrete Data Refinement

Following the specified concept, a prototype has been implemented for tree stem data processing in the forestry domain (the data refinement flow is illustrated in Fig. 4). There will be a data item collection for each processed tree stem and the logs made from it. First, measurement values are retrieved and structured as data item collections. Then, an outlier check is performed for each measurement value in each data item collection; the data items that do not match their conditions are marked as failed. Next, appearance items are resolved by checking whether each data item collection satisfies each appearance condition set or not. Finally, stem data item collections are categorised based on their values. Here, it must be noted that if some measurement value required for categorisation has failed (per outlier check), the category cannot be resolved. Instead, the stem data item collection (and the related log data item collections) will not be further processed.

The method utilised for the outlier check is straightforward. For each measurement, an arbitrary number of conditions may be specified. In a typical case, there will be a lower and an upper bound. While the utilised outlier detection method is simple, various more advanced methods exist as discussed by Hodge and Austin [11], for example. An XML (Extensible Markup Language) format has been designed to have configurable outlier conditions for each data item.

To enable configurability, the conditions for appearance items are defined with the same XML format as the outlier limits. For each appearance item, an arbitrary set of data items may be inspected. For each data item, there can be an arbitrary number of conditions (similar to each data item that may have multiple outlier conditions).

While various context recognition methods exist, the prototype utilises a simple though configurable way. The prevailing context is determined by finding the most typical stem data item collection category. That category is considered the context; any other data item collections are excluded from further processing as they are considered exceptions in the current environment. Categories are defined using a tree-like condition set (see Fig. 5): the categorisation tree may inspect any data items to resolve the category of a data item collection. The categorisation tree is stored in a structured text document generated in a fleet-wide data analysis. The prototype parses the categorisation tree so it is available in the application during machine operation. Similar to outlier and appearance condition definitions, even the categorisation tree is transferred as a configuration item to each machine.

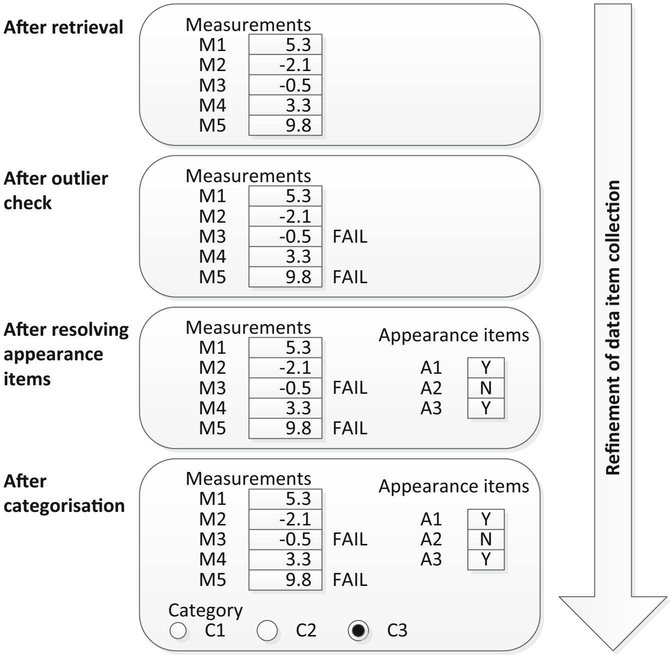


Fig. 4. Data refinement in the prototype implementation [15].

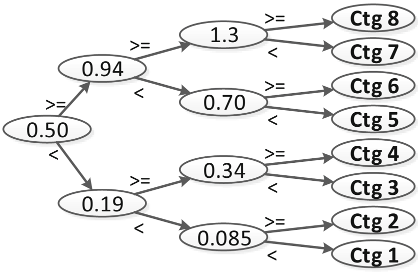


Fig. 5. An example of categorising a tree stem after its volume in m^3 (though there could be multiple variables observed in the conditions). Here, the categories have indices from 1 to 8, a high index indicating a large stem. For instance, category 4 has the stems with a volume within range [0.34–0.50] [15].

5.2 Software Implementation

Figure 6 illustrates the concrete software implementation of the prototype. The prototype may be roughly divided to a cloud side and a machine side; both the sides are explained in more detail in the following paragraphs.

The cloud side covers machine learning functions as well as data refinement configuration access points. The utilised cloud environment is Microsoft Azure. In the prototype, no machine learning is performed in the cloud as it is out of the

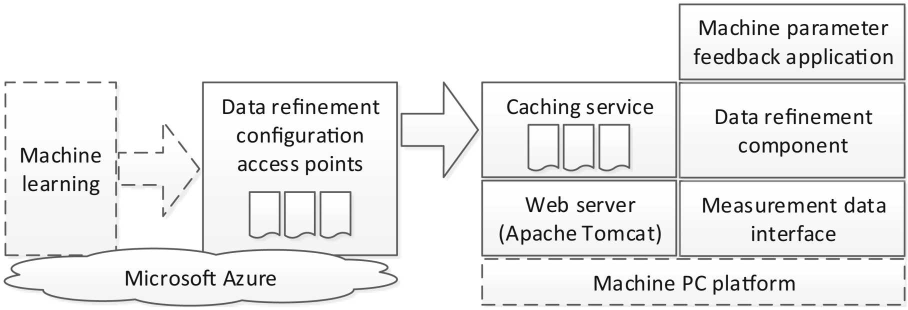

Machine learning
Data refinement configuration access points
Microsoft Azure
Caching service
Web server (Apache Tomcat)
Machine parameter feedback application
Data refinement component
Measurement data interface
Machine PC platform

Fig. 6. Prototype architecture.

scope of this study. Still, as Azure has the facilities to store large amounts of data and even machine learning capabilities, it is considered an appropriate platform. All the configuration items (the conditions for measurement outliers, appearance items and tree stem categorisation) are located in Azure to demonstrate their accessibility from a HTTP-based REST API in the cloud.

Due to non-persistent internet connectivity, a caching web service has been implemented to provide an access to configuration items locally in machinery. The web service has been implemented with Java and it is run on a Tomcat web server in a desktop computer. Modern machinery often run their equipment and operation related software on a PC platform, which makes it possible to install a general-purpose web server even there.

The actual configurable data refinement component has been utilised in an application that assists the machine operator to optimise various equipment parameters during machine operation. Although run in a desktop PC, the execution environment is realistic as a measurement data interface identical to a physical machine is utilised; besides, the interface has been set up to provide data collected during actual physical machine operation.

The classes of the data refinement component prototype are illustrated in Fig. 7. An abstraction called *item condition* is essential: it defines a condition for a data item (such as a measurement). Item conditions are utilised for both outlier checking and specifying appearance items. Each item condition is a part of an *item condition definition* (as a value may have multiple boundaries), and each item condition definition is a part of an *item condition definition set* (such as the conditions of an appearance item). Item conditions are stored in an XML configuration file parsed by the *item condition XML reader* class. *Appearance resolver* class resolves which appearances are true for each data item collection. The conditions for data item collection categorisation are parsed by the *categorisation tree parser* class.

The data refinement component has been implemented with Java although any other platform could be used as well. As long as component interfaces (such as configuration formats) are as specified, even heterogeneous platforms are possible within an enterprise.

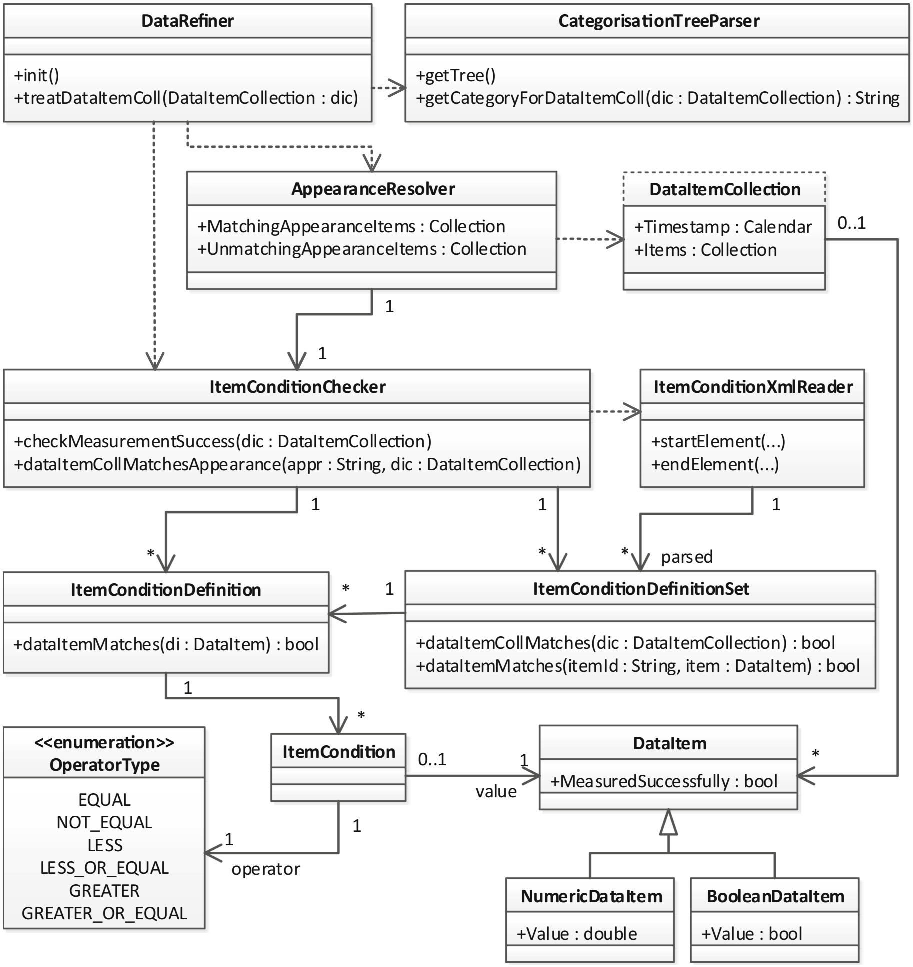

DataRefiner
+init()
+treatDataItemColl(DataItemCollection : dic)
CategorisationTreeParser
+getTree()
+getCategoryForDataItemColl(dic : DataItemCollection) : String
AppearanceResolver
+MatchingAppearanceItems : Collection
+UnmatchingAppearanceItems : Collection
DataItemCollection
+Timestamp : Calendar
+Items : Collection
0..1
1
1
ItemConditionChecker
+checkMeasurementSuccess(dic : DataItemCollection)
+dataItemCollMatchesAppearance(appr : String, dic : DataItemCollection)
ItemConditionXmlReader
+startElement(...)
+endElement(...)
1
1
1
*
*
*
parsed
ItemConditionDefinition
+dataItemMatches(di : DataItem) : bool
*
1
ItemConditionDefinitionSet
+dataItemCollMatches(dic : DataItemCollection) : bool
+dataItemMatches(itemId : String, item : DataItem) : bool
1
*
<<enumeration>>
OperatorType
EQUAL
NOT_EQUAL
LESS
LESS_OR_EQUAL
GREATER
GREATER_OR_EQUAL
ItemCondition
0..1
1
value
DataItem
+MeasuredSuccessfully : bool
*
1
1
operator
NumericDataItem
+Value : double
BooleanDataItem
+Value : bool

Fig. 7. The classes of the data refinement component prototype. Adapted from [15].

5.3 Practical Experiment with Machine Data

The prototype has been utilised in the refinement of real operational forestry data in a machine parameter optimisation application. The application estimates machine performance and suggests parameter tuning in case the parameters seem non-optimal. As the number of machine parameters may reach hundreds in a modern machine, their optimisation is difficult for a typical operator. That is, such information refinement has considerable added value to the operating enterprise. The actual parameter optimisation application utilises the outcome of the data preprocessing introduced in this paper. As real operating data and realistic interfaces are utilised, the setup is almost identical as if the application were run in the field. Kannisto et al. [16] have already considered the scenario with parameters rather than data preprocessing in the scope.

Parameter optimisation is not a simple task as it requires multiple factors to be considered. The operating context and the type of work being performed may affect both which parameter values result in a good performance and the actual performance values. Large amounts of historical data should be analysed to generate reference sets of performance values and optimal parameter values. As machines keep operating, data should be continuously collected to refresh parameter related knowledge; as knowledge updates are delivered to multiple machines, ease in management becomes beneficial. Knowledge generation actions require both extensive domain expertise and advanced data refinement methods so they should be performed by a dedicated group of skilled personnel. The knowledge may be managed by, for instance, the machine manufacturer or a fleet operator.

In this demonstration, the function under parameter optimisation is automatic tree stem positioning in a wood processing implement. Stems are positioned to be cut into logs. Such a case suits well for parameter optimisation as automatic positioning is controlled entirely by machine parameters rather than by the operator – the most of other machine functions are largely affected by operator skills.

The outliers of two measurements are observed in the experiment. *Positioning error* describes how close to its optimal cutting position a stem has been stopped. In contrast, *feed speed* does not determine positioning performance but it is an important measure as the overall machine performance is estimated in further data processing (more speed results in a higher productivity value). The outlier conditions are as follows: feed speed cannot be negative, and the absolute value of positioning error must be within 30 cm of the desired position.

Stem categorisation is important in the experiment. According to stem volume, each stem is put into one of eight categories. As little trees are not of interest in this felling scenario, there is an additional condition that each stem with a felling diameter of less than 15 cm is excluded. The context recognition method also uses the outcome of the categorisation. It is simplistic: for each category, there is a directly mapped context class. The stems in any other category are considered irrelevant and excluded from further processing.

In the experiment, appearance items have an informative function. They are generated using conditions that specify if a stem represents a long spruce or a long pine (that is, both tree species and stem length are observed). For the resulting Boolean *true* values, percentages are calculated how large their section is within the relevant stem category (or context; e.g. "64% of stems are long spruces"). While the parameter analysis application does not utilise these percentage values, the machine operator might want to observe themselves if tree species or lengths actually affect optimal parameter values. If there are such factors, they should actually be discovered in fleet level data analyses. Then, they could be utilised by the parameter optimisation application in the field. From the conceptual point of view, the configurable indirect variable calculation feature improves management possibilities in data preprocessing.

6 Results and Discussion

The objective of this work was to design a software concept to enable the centralised management of data refinement in an arbitrarily large geographically distributed machine fleet. Outlier inspection for measurements was required as well as data set categorisation and the possibility to specify variables calculated from original data. Context recognition and consideration were also required.

The concept meets its information management requirements well. The ease of management of the application workflow was considered paramount: it is possible to configure not only outlier limits but also data set categorisation and the context recognition method. In addition, it is possible to specify variables for information inferred from explicit measurement data. Such data may be numeric (calculated) or Boolean values (resulting from the assertion of multiple conditions). The concept enables data collection from machines, machine learning to generate the configuration items as well as access to the configuration items managed in a cloud environment.

A functional cloud-managed data refinement software component prototype has been implemented. It implements the specified data refinement flow. First, an outlier check is performed on measurement values followed by the calculation of derived variables. Then, each data item collection (a data set of key-value pairs) is categorised according to specified conditions, and finally, the prevailing context is determined using categorisation information. The configurability requirement is fulfilled by getting outlier conditions, derived variable calculation conditions and categorisation definition from a cloud service. Machine mobility and geographic distribution have also been considered by implementing a caching service run locally in each machine.

The concept has been experimented with real operative data from 11 forestry machines. For each machine, the data of thousands of stems was processed so there has been a lot of repetition in application cycles. The outcome of the software component (i.e. refined data) was utilised to optimise the parameters of automatic tree stem positioning in a wood processing implement. The data refinement results are in Table 3. In each data set, the number of stems in the context was relatively low. The context recognition method returned the same operating context for each data set (stems with volume within 0.19–0.34 m^3) so it is not included in the table.

The outlier results provided by the component seem useful. For positioning error values, the exclusion percentage is relatively low – mostly less than 1%, at most 1.4%. However, the highest exclusion percentage due to feed speed value is 9.7%. If these values were not excluded from further processing, they could cause significant errors in further calculations performed by other applications. Still, depending on error magnitudes, even a 1% section of erroneous values may cause misleading results.

18–54% of all stems were excluded from further processing as their felling diameter was less than 15 cm. The percentages are relatively high. As the parameter optimisation goal was concerned with the processing of large stems, such large amounts of relatively little stems might distort further calculations.

Table 3. Data refinement results with real forestry machine operation data [15]

Mach ID	Stems	Logs	Feed sp. outlier (logs)	Pos. error outlier (logs)	Stems excluded (felling diam <15 cm)	Stems in context	Long spruces (context)	Long pines (context)
1	11, 000	27,000	4.0%	0.33%	54%	1,400	40%	52%
2	6, 300	19,000	1.8%	1.1%	23%	1,200	60%	26%
3	14, 000	39,000	4.1%	0.93%	36%	2,500	61%	22%
4	6, 600	18,000	3.9%	0.56%	48%	1,100	61%	5.6%
5	5, 900	18,000	2.9%	0.27%	31%	1,000	60%	8.7%
6	7, 800	26,000	5.1%	0.36%	30%	1,100	75%	9.1%
7	8, 000	27,000	1.6%	0.39%	26%	1,400	72%	7.9%
8	10, 000	28,000	4.9%	0.76%	27%	2,000	33%	33%
9	12, 000	38,000	4.9%	1.4%	34%	1,600	64%	20%
10	6, 800	25,000	9.7%	0.93%	18%	1,100	55%	4.2%
11	6, 500	20,000	4.9%	1.0%	29%	1,400	62%	13%

However, it may also be asked if the processing of little stems should also be considered in optimisation. In that case, their data should be passed through distinguished from large stems.

The percentages of long spruces and pines are also included in the results table. In most cases, spruce appears the dominant species. The parameter analysis application did not utilise this information for anything so it is purely informative in the experiment.

The context recognition method appeared to be ineffective as its result was the same context class for each test run. More context recognition and classification related research should be performed. The goal of context recognition should be reconsidered; that would specify which variables and what kind of methods should actually be included as the context is determined. However, the task is more related to domain expertise and data analysis rather than the knowledge management concept relevant in this study. In the end, it might be beneficial if the entire context recognition method could be updated along with the configuration.

The experiments made with the prototype indicate that the data refinement management concept is functional and valuable. It has potential business value in real-life data processing: it would be easier to manage the refinement of the data consumed by various end user applications. Such applications may, for instance, assist in more effective machine operation. However, the prototype also has room for further development. Context recognition should be studied further to provide more practical value. Derived variables can only be Boolean values – numeric values are not currently supported though they would offer significantly more potential for various uses cases. In addition, even though configuration documents are already managed with cloud services, their coupling with concrete machine learning methods in the cloud are not covered. The prototype should be developed further to cover the entire chain of data collection, data storage and machine learning chain. While data analysis may be applied in any environment, a cloud promotes scalability and availability, which is beneficial for a large enterprises in a distributed environment. Ultimately, it would be interesting to see the concept in operation in an everyday business environment. Finally, a long-term development need is the consideration of individual machine characteristics. In practice, individual differences may affect machine performance and sensor readings – this stems from, for instance, individual hydraulic component characteristics or the degree of abrasion. This also affects how raw sensor data should be preprocessed. In the future, machine learning should be applied *locally* in each machine to consider such differences.

7 Conclusion

In this study, a software system concept is introduced to enable centralised management for measurement data refinement within a distributed machine fleet. Modern machines have been equipped with advanced ICT devices that enable added-value software for various purposes (such as operator feedback for more efficient operation). To ease application development, the data refinement concept covers configurability for multiple important data preprocessing tasks including outlier detection, the calculation of derived variables and context recognition. From the management point of view, the concept covers measurement data collection, the utilisation of machine learning methods to generate data refinement configurations as well as configuration item access points – all in a cloud.

Following the concept, a functional data refinement management prototype has been implemented. It is an intermediary component that refines measurement data using configuration items received from cloud services. The prototype has been executed as a part of an application that provides assistance in machine parametrisation. Experiments with real operational measurement data have demonstrated the practical value of the concept: how machine data refinement management can be largely facilitated with cloud services.

There are also future research tasks. While successful, the prototype should be developed further to meet all the requirements of the concept. Also, the concept should cover even machine learning run locally in machines to consider individual machine characteristics.

Acknowledgments. This work was made as a part of the D2I (Data to Intelligence) project funded by Tekes (the Finnish Funding Agency for Innovation). The authors would like to express their sincere gratitude to the project partners and participant companies.

References

1. Bahga, A., Madisetti, V.K.: Analyzing massive machine maintenance data in a computing cloud. IEEE Trans. Parallel Distrib. Syst. **23**(10), 1831–1843 (2012). https://doi.org/10.1109/TPDS.2011.306
2. Banerjee, T.P., Das, S.: Multi-sensor data fusion using support vector machine for motor fault detection. Inf. Sci. **217**, 96–107 (2012). https://doi.org/10.1016/j.ins.2012.06.016

3. Basir, O., Yuan, X.: Engine fault diagnosis based on multi-sensor information fusion using Dempster-Shafer evidence theory. Inf. Fusion **8**(4), 379–386 (2007). https://doi.org/10.1016/j.inffus.2005.07.003
4. Choudhury, T., et al.: The mobile sensing platform: an embedded activity recognition system. IEEE Pervasive Comput. **7**(2), 32–41 (2008). https://doi.org/10.1109/MPRV.2008.39
5. Deng, L., Yu, D.: Deep learning: methods and applications. Found. Trends Sig. Process. **7**(34), 197–387 (2014). https://doi.org/10.1561/2000000039
6. Duan, L., Xu, L.D.: Business intelligence for enterprise systems: a survey. IEEE Trans. Ind. Inform. **8**(3), 679–687 (2012). https://doi.org/10.1109/TII.2012.2188804
7. Favela, J., et al.: Activity recognition for context-aware hospital applications: issues and opportunities for the deployment of pervasive networks. Mob. Netw. Appl. **12**(2–3), 155–171 (2007). https://doi.org/10.1007/s11036-007-0013-5
8. Filev, D., Lu, J., Hrovat, D.: Future mobility: integrating vehicle control with cloud computing. Mech. Eng. **135**(3), S18–S24 (2013)
9. Fountas, S., Sorensen, C., Tsiropoulos, Z., Cavalaris, C., Liakos, V., Gemtos, T.: Farm machinery management information system. Comput. Electron. Agric. **110**, 131–138 (2015). https://doi.org/10.1016/j.compag.2014.11.011
10. Golparvar-Fard, M., Heydarian, A., Niebles, J.C.: Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. Adv. Eng. Inform. **27**(4), 652–663 (2013). https://doi.org/10.1016/j.aei.2013.09.001
11. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. **22**(2), 85–126 (2004). https://doi.org/10.1007/s10462-004-4304-y
12. Hou, L., Bergmann, N.: Novel industrial wireless sensor networks for machine condition monitoring and fault diagnosis. IEEE Trans. Instrum. Meas. **61**(10), 2787–2798 (2012). https://doi.org/10.1109/TIM.2012.2200817
13. Iftikhar, N., Pedersen, T.B.: Flexible exchange of farming device data. Comput. Electron. Agric. **75**(1), 52–63 (2011). https://doi.org/10.1016/j.compag.2010.09.010
14. Jardine, A.K., Lin, D., Banjevic, D.: A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Sig. Process. **20**(7), 1483–1510 (2006). https://doi.org/10.1016/j.ymssp.2005.09.012
15. Kannisto, P., Hästbacka, D.: Enabling centralised management of local sensor data refinement in machine fleets. In: Proceedings of the 8th International Conference on Knowledge Management and Information Sharing, vol. 3, pp. 21–30 (2016). https://doi.org/10.5220/0006045600210030
16. Kannisto, P., Hästbacka, D., Kuikka, S.: System architecture for mastering machine parameter optimisation. Comput. Ind. **85**, 39–47 (2017). https://doi.org/10.1016/j.compind.2016.12.006
17. Kannisto, P., Hästbacka, D., Palmroth, L., Kuikka, S.: Distributed knowledge management architecture and rule based reasoning for mobile machine operator performance assessment. In: Proceedings of the 16th International Conference on Enterprise Information Systems, pp. 440–449 (2014). https://doi.org/10.5220/0004870004400449
18. Khot, L.R., Tang, L., Blackmore, S., Nørremark, M.: Navigational context recognition for an autonomous robot in a simulated tree plantation. Trans. ASABE **49**(5), 1579–1588 (2006)

19. LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big data, analytics and the path from insights to value. MIT Sloan Manag. Rev. **52**(2), 21–31 (2011)
20. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1038/nature14539
21. Lu, B., Gungor, V.: Online and remote motor energy monitoring and fault diagnostics using wireless sensor networks. IEEE Trans. Ind. Electron. **56**(11), 4651–4659 (2009). https://doi.org/10.1109/TIE.2009.2028349
22. March, S.T., Smith, G.F.: Design and natural science research on information technology. Decis. Support. Syst. **15**(4), 251–266 (1995). https://doi.org/10.1016/0167-9236(94)00041-2
23. Osborne, J.W., Overbay, A.: The power of outliers (and why researchers should always check for them). Pract. Assess. Res. Eval. **9**(6), 1–12 (2004)
24. Palmroth, L.: Performance monitoring and operator assistance systems in mobile machines. Ph.D. thesis, Department of Automation Science and Engineering, Tampere University of Technology, Tampere, Finland (2011)
25. Peets, S., Mouazen, A.M., Blackburn, K., Kuang, B., Wiebensohn, J.: Methods and procedures for automatic collection and management of data acquired from on-the-go sensors with application to on-the-go soil sensors. Comput. Electron. Agric. **81**, 104–112 (2012). https://doi.org/10.1016/j.compag.2011.11.011
26. Steinberger, G., Rothmund, M., Auernhammer, H.: Mobile farm equipment as a data source in an agricultural service architecture. Comput. Electron. Agric. **65**(2), 238–246 (2009). https://doi.org/10.1016/j.compag.2008.10.005
27. Stiefmeier, T., Roggen, D., Ogris, G., Lukowicz, P., Tröster, G.: Wearable activity tracking in car manufacturing. IEEE Pervasive Comput. **7**(2), 42–50 (2008). https://doi.org/10.1109/MPRV.2008.40
28. Tao, F., Zhang, L., Liu, Y., Cheng, Y., Wang, L., Xu, X.: Manufacturing service management in cloud manufacturing: overview and future research directions. J. Manuf. Sci. Eng. **137**(4), 040912 (2015). https://doi.org/10.1115/1.4030510
29. Väyrynen, T., Peltokangas, S., Anttila, E., Vilkko, M.: Data-driven approach for analysis of performance indices in mobile work machines. In: Data Analytics 2015, The Fourth International Conference on Data Analytics, pp. 81–86 (2015)
30. Wan, J., Zhang, D., Zhao, S., Yang, L.T., Lloret, J.: Context-aware vehicular cyber-physical systems with cloud support: architecture, challenges, and solutions. IEEE Commun. Mag. **52**(8), 106–113 (2014). https://doi.org/10.1109/MCOM.2014.6871677
31. Whaiduzzaman, M., Sookhak, M., Gani, A., Buyya, R.: A survey on vehicular cloud computing. J. Netw. Comput. Appl. **40**, 325–344 (2014). https://doi.org/10.1016/j.jnca.2013.08.004
32. Wu, D., Rosen, D.W., Wang, L., Schaefer, D.: Cloud-based design and manufacturing: a new paradigm in digital manufacturing and design innovation. Comput.-Aided Des. **59**, 1–14 (2015). https://doi.org/10.1016/j.cad.2014.07.006
33. Yang, B.S., Kim, K.J.: Application of Dempster-Shafer theory in fault diagnosis of induction motors using vibration and current signals. Mech. Syst. Signal Process. **20**(2), 403–420 (2006). https://doi.org/10.1016/j.ymssp.2004.10.010

Knowledge Management in Enterprise Architecture Projects

Juan Pablo Meneses-Ortegón(✉) and Rafael A. Gonzalez(✉)

Pontificia Universidad Javeriana, Bogotá, Colombia
{juan.meneses, ragonzalez}@javeriana.edu.co

Abstract. In the domain of enterprise architecture (EA), project and team complexity can be a challenge as well as an opportunity for knowledge management (KM). EA projects generate a series of artifacts that contain knowledge directly or indirectly which can be reused or transferred from project to project. In this paper, the interest in providing a KM framework for TOGAF-based EA, to capture, store and reuse lessons learned in the first phases of the project. The framework is described in a meta-model known as "ways of", addressing the ways of thinking, working, supporting, controlling and modeling. Validation is presented through a case study in a consulting company and through expert opinion.

Keywords: Knowledge management · Enterprise architecture
TOGAF · Preliminary · Architecture vision

1 Introduction

"Enterprise Architecture" - EA is a discipline that is defined as "a coherent set of principles, methods and models used in the design and/or implementation of an organizational structure, business processes, information systems, and infrastructure" [1]. This discipline involves and requires the effective use of both tacit and explicit knowledge, related to both client and consulting companies. Due to the complexity of this knowledge, companies need a flexible processes that allows them to adapt themselves as such knowledge evolves [2].

Of existing EA frameworks, "The Open Group Architecture Framework" - TOGAF stands out, due to its world-wide acceptance and use. This framework proposes several phases to follow, including two early stages: preliminary and phase A (architecture vision). These phases provide the initial knowledge that allows supporting the rest of the enterprise architecture exercise, requiring crucial knowledge management processes, such as identification, acquisition, and development [3]. Moreover, resulting knowledge in EA consulting firms may become their most valuable resource.

TOGAF has an associated lifecycle to develop the enterprise architecture called "Architecture Development Method" - ADM [4], that presents specific steps that generate information which can be converted into knowledge in order to be used by the client as well as the consulting firm. When a company doesn't have a model or policy for knowledge management, its knowledge can be lost or not effectively re-used.

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 287–305, 2019.
https://doi.org/10.1007/978-3-319-99701-8_14

This article provides a research proposal, based on knowledge management, to support communication, transmission and appropriate use of knowledge for decision-making in EA projects. It gives support to the patterns of enterprise architecture management, such as the definition of methodologies, visualization and representation of information models [5]. Thus, by using management services centered on explicit knowledge generated through ADM within the TOGAF framework and stored in the architecture repository, it allows improved governance of the implementation process.

This paper is a revised and extended version of an early presentation of a proposed framework [6] for enterprise architecture knowledge management, which is added a theoretical approach to support the knowledge transfer in architecture projects. This paper has seven sections that describe the research carried out. Initially, the topics used for the research as such as knowledge, knowledge management and enterprise architecture are conceptualized. Then, the research problem, the methodology used, followed by the presentation of the case study used in a specific application domain. In the next section, the article shows the proposed knowledge management framework and its composition. Finally conclusions and future work derived from the project are presented.

2 Conceptual Framework

In order to define a relationship between Enterprise Architecture and Knowledge Management it is necessary to analyze some concepts, classifications, properties and application frameworks. Based on this, the correlation between EA and KM is presented in the research approach section.

2.1 Knowledge

All human activities create a large amount of data and information, this increases the knowledge value and its use as a strategy [7], that’s way KM is “rapidly becoming an integral business activity for organizations as they realise that competitiveness pivots around the effective management of knowledge” [8]. It is necessary to differentiate the data of the information and the information of the knowledge. Data is any number, word, e-mail, etc.. When these data have a specific role in a context, the same context creates a relationship between them, this is information. And when someone or something does an analysis about this information in order to determinate something about the organization, this is knowledge [7]. Related definitions, include [9], where it is a “a mind state, an object, a process, a condition of having access to the information”. In [10] it is defined as a mix between experience, value, contextual information and experts vision that bring new knowledge and innovation to an organization.

Types of Knowledge

Typically knowledge is classified as tacit or explicit (that is in documents, e-mails, manuals, videos, etc.) [7, 9]. For [10] in an organization there are professional knowledge (which correspond to a specific functional domain) [11] and firm-specific knowledge (which is hard to replicate because it’s related to specific products or services) [11]. In any case, any type of knowledge has to be well used in order to give the organization a competitive advantage [7].

2.2 Knowledge Management

Knowledge management – KM is the process of applying a systematic approach to capture, structure and manage knowledge throughout the organization in order to work quick, reuse the best practices and reduce the expensive reset of a project [12]. KM includes strategies or process for identifying, capturing and exploiting knowledge" [13].

When KM is in place it often follows a cyclical approach, involving a set of activities, such as: "creation, transfer and application", "capture, transfer and application" and "identification, capture, development, diffusion, application and storage" [13].

Knowledge Management Framework

Generally, knowledge management frameworks - KMF are developed in research and are classified as: prescriptive, descriptive or both. Prescriptive frameworks formulate the activities to manage knowledge. Descriptive frameworks start from gathering data and abstracting the description of the most important attributes that contribute to the success of a knowledge management process.

The KMF focus is in the management activities and must be integrated with the organization's goals and strategies, and with the people who intervene in the management process. Also, KMF must increase an organizational culture oriented towards knowledge [14].

2.3 Enterprise Architecture

Enterprise architecture - EA is a discipline that looks for the alignment between processes and IT resources in an enterprise [15]. [1] defines enterprise architecture as a coherent set of principles, methods and models which are used in the design and implementation of an organizational structure of the enterprise, its business processes, information systems and infrastructure.

Enterprise architecture has emerged from technology advances and their role in supporting and transforming business processes. It helps to integrate the IT tools and the business in an organization. This integration develops IT criteria according to the organization's mission and business strategies, processes and activities [16].

Within the frameworks to the implementation of EA stand out [1]: Zachman Framework (created by Jhon Zachman), TOGAF (by Open Group) and Model Driven Architecture – MDA (by Object Management Group – OMG). This papers is focused on TOGAF, describes as follows.

TOGAF

This framework, created by The Open Group [16], develops, designs and implements an EA. TOGAF is supported by a model that proposes some phases for an architecture project based on a set of interactions and elements to help with the architecture's maintenance and management. Within these common support elements, one is called "enterprise continuum" which supports the architecture based on tools and models. In this way, architectural and business information gets integrated in a repository [16].

TOGAF proposes a model to the architecture's life cycle called Architecture Development Model - ADM. This model works phases to follow in each interaction of an EA project. This project is based on the two first phases: preliminaries and architecture vision.

Preliminaries

It is a phase to stablish the bases to start the EA project. In this phase the people team defines *where, what, why, how* and *who* is going to do the architecture. This phase works with the enterprise understanding to propose methodologies to be used in the development and tools to its support [16].

Architecture Vision

This is the first phase of ADM (the preliminaries is before start the phases) and in here the team works in activities as: review architecture principles and business principles, the definition of the goals and scope of the project, the stakeholders' identification and the statement of architecture work and secure approval [16].

2.4 Knowledge Management and Enterprise Architecture

At this point it is necessary to identify some relation between KM and EA with some projects where have worked both areas at the same time.

As we mentioned in the last section, EA defines the way IT is designed and implemented in an organization, but it does not take into account that EA has to deal with human behavior included in new roles, duties and responsibilities, where KM is needed [17]. Organizations that define an EA require organizational learning in order for their members to have access to it. The management of this knowledge considers factors such as acquisition, diffusion and storage of knowledge for the organization [18].

[19] use different tools to manage knowledge, they work with EA and business intelligence to present a KM framework. In [20] the management of an EA in a KM context, through the actual against planned components of the architecture, which are stored and maintained. In this work KM is used for EA to have a better way of organizing enterprise knowledge.

Another example of the use of KM in enterprise architecture is the development of a collaborative platform [21]. The authors suggest a KM process which is realetd to the production of a specific EA.

Lastly, the use of datawarehousing technologies within an EA repository is presented in [22] as actives for a knowledge management framework.

3 Research Approach

This section presents the research problem and the methodology used to face it.

3.1 Research Problem

TOGAF enterprise architecture is supported in the ADM method [4]. In each phase, a number of deliverables and associated knowledge is generated, but it may become lost or not effectively reused, due to lack of monitoring in consulting firms. This knowledge is important because of to the possibility of taking advantage of it in later enterprise architecture projects. Information that becomes knowledge, by being linked to experience, will help to deal with future projects, allowing the project manager to not repeat the same mistakes or indeed to take advantage of good practices identified in previous projects. These good practices will guide the development of new activities, for instance, in activities applied to government-related projects or within the same industry sector. In the same way, the KM has some initiatives in order to encourage the organization to not lose staff [23]. When this kind of people leaves, the organization can lose knowledge as lessons learned or good job practices they carried out inside the development of projects and applications.

The initial phases of TOGAF are a particularly rich source of potentially valuable knowledge. In the preliminary phase of TOGAF-ADM, the EA group defines the project's goals and expectations according to the aims and vision of the business and the definition of stakeholders, their requirements and priorities. All this implies a process of knowledge identification and acquisition related to frameworks, methodologies and other tools that support the rest of the project. After defining this initial stage, the architecture vision goes on to further specify knowledge from the point of view of business, data, application and technology [3]. However, there is no evidence of effective future use of these outputs in future projects in a governed and systematic fashion, attached to ADM.

Although enterprise architecture is often supported in knowledge management tools through the implementation of enterprise wikis or digital libraries that allow information retrieval [24, 25], and other kind of project-oriented search-based tools for enterprise architecture management [26], EA has not been sufficiently supported in tailored knowledge management processes.

The contribution of this paper is aimed at communicating and/or transferring knowledge for EA decision-making. This decision-making may be reflected, for instance, in activities such as reviewing the current architecture [27] or in supporting the patterns of enterprise architecture management, like methodologies definition, visualization and representation of information models [5]. This flow of knowledge is a special challenge because of the number and diversity of stakeholders.

Likewise, it is important to manage the generated artifacts as an important part of the enterprise architecture, which are usually stored in the architecture's repository. The problem is that often these repositories are no more than that, a repository where the results of each activity are stored, but do not inform future decisions based on reuse or socialization. Therefore, the use of explicit knowledge management services integrated in the process of ADM - TOGAF, coupled to the repository, should allow the implementation of an effectively governed architecture where the knowledge acquired is exploited beyond the scope of a single project.

3.2 Research Methodology

The research methodology adopted is design science research as a methodology for the design and development of information systems [28], where the designed artefact in our case is the framework for EA knowledge management. In Peffers et al., the process is effectively completed once the artefact is demonstrated, validated and communicated.

For this reason, the development of this project was implemented through this methodology, which focuses on solving real-world problems through a 6-phase approach:

- Identify the problem and motivation
- Define solution objectives
- Design and development
- Demonstration
- Evaluation
- Communication.

Development and validation of the solution was done iteratively around the following artefacts: knowledge maps, knowledge processes, process models of the preliminary and architecture vision stages and the overall knowledge management framework.

Specific development of the research method is described in Sects. 4 and 5.

3.3 Case Study

Development of the knowledge management framework was carried out in the context of a large IT and EA Colombian consulting firm, Indra Colombia, in relation to their TOGAF-based EA consulting projects. Indra is a multinational company with headquarters in Spain and its main core is generating innovative IT services and solutions. These services are delivered in line with management strategies of customer needs through consulting, development and project management, integration and implementation solutions and outsourcing of information systems, in sectors such as: transport and traffic, energy and industry, public administration and healthcare, financial services, security and defence and telecom and media [29].

Indra has offices in 138 countries with a total of 42 thousand professionals approximately. The company has presence in Latin American countries including Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, El Salvador, México, Panamá, Perú, Uruguay and Venezuela [29]. Indra has an 18 year presence in the Colombian market, currently with more than 2000 professionals and 7 offices in Bogotá, Pereira, Barranquilla and Medellin, with solutions and services in cloud computing, outsourcing of BPO (business process outsourcing) and networks and telecommunications, with major clients in the public and private sector. Our scope is focused on EA consulting, which has mostly been oriented to the financial, healthcare and public sectors.

Our first step was getting to know their EA processes. To describe those processes, we gathered information through meetings (1 h each) with the consulting area manager, leaving as evidence the minutes of each one of them. In these meetings, we uncovered their enterprise architecture processes already undertaken through documents resulting from projects and proposals previously made by the consulting area.

After having the information of the company processes, we matched them against the activities proposed by TOGAF, in order to identify which tasks were completely carried out, which ones were not and which ones could be most amenable for knowledge management.

Subsequently, we analysed documents (proposals and enterprise architecture artefacts), complemented with informal interviews and direct observation of the activities carried out to fully address explicit and tacit knowledge considerations.

4 Proposed Framework

In this section, we describe how we designed the knowledge management framework (KMF).

4.1 Why the Framework

The main task for the design of a KMF is the definition of its purpose.

The initial premise of this research is to enable the use and reuse of knowledge. This was motivated by the aim of speeding up the development of the initial phases of an EA project in order to generate knowledge to provide innovation in new proposals of projects (to reuse knowledge and improve upon it). After that, we wanted to manage processes of knowledge generation and its storage. To do this, the framework was instantiated in a prototype, allowing its use and validation. With this prototype, the extraction and dissemination of knowledge were enhanced. The prototype was structured around the early stages of TOGAF trying to maintain or improve delivery times. Indeed, it is important to note that while reuse, per se, is often a time-saving strategy, knowledge management activities that enable such reuse may, by contrast, take up a significant amount of time, which is partly the reason why in practice it is not often found to a large extent.

To meet these goals, first we seek to identify the elements that generate information, through use of ontological engineering and other knowledge management methods, abstracted using the "Ways of" meta-model [30]. In this meta-model, we identified the tasks to develop both the knowledge management framework as well as the prototype, the way these tasks are modelled, the languages to be used for development and how to control the outcome. The complete "Ways of" meta-model guiding our framework is presented in Sect. 4 of this document.

4.2 Knowledge – Generating Entities

The description of the entities associated with knowledge and learning processes were based on the case study. In this description, we identified two components: the first begins with personal interaction to gather information from face to face meetings, which may be with the consulting firm or the client company. The second was focused in the acquisition and storage of the artifacts generated by each activity. In Table 1 we describe the spaces or objects used for information generation, which is associated with the enterprise architecture in the selected phases (see Table 1).

Table 1. Enterprise architecture information [6].

Object	Description
Personal relationship	For the development of the early stages of TOGAF, people involved create efficient communication, but knowledge is often in conversations and reuse is not possible
E-mail	E-mail as a tool used to obtain information from customer or to exchange information among those involved in the project
Previous proposals	We take information from development of proposals already made that were approved or not
Success cases	Among the projects already developed, it is important to identify success stories that can provide feedback to be used in future projects. This will include artefacts generated in previous projects, and unrealized (projects in which the company made proposal but were not developed, yet contain useful information)
Lessons learned	In each project proposal, which was developed or not, the project generates some lessons learned in order not to make the same mistakes, if any

4.3 Meta-model Framework

To develop the framework, we selected the “Ways of” meta-model because this is an appropriate way of abstracting the results of an EA processes, according to [30]. With this meta-model, used as a template, we described how the framework was generated, how it should be used and how it will be supported by IT elements.

Based on this model, we start by describing the way of thinking, which shows the understanding of the domain in which the framework will be applied in relation to the issues raised. This way helps to understand how processes can be modelled and to take a broad view of the solution.

We also include the way of modeling, which identifies how a process is modeled and what language is used for it, the activities and tasks of the framework, as well as identifying the relationship between them.

The way of working, is the next step, it describes what tasks are performed in the framework and their order.

The way of controlling, indicates the tools that enable monitoring how the framework objectives are being fulfilled, based on the use of resources.

Finally, the way of supporting determines the IT that will be used to support the tasks and/or activities in the framework.

The design of this knowledge management framework focused on the first two phases of TOGAF-based enterprise architecture, known as preliminary and Phase A. Vision Architecture.

These stages were chosen because they are sequential and are the initial phases of ADM. This allows developing a knowledge management process of an enterprise architecture project since its inception and with often more reusable content than later stages.

5 A Knowledge Management Framework for Enterprise Architecture

In this segment, we describe the five "Ways of" that constitute the framework: thinking, modelling, working, controlling and supporting.

5.1 Way of Thinking

The "way of thinking" is designed for the first phases of TOGAF, preliminary and architecture vision. The aim in the first one (preliminary) is to build the bases needed in order to start the enterprise architecture project, that is, in this phase we define the "where, why, how and who" to build the architecture. The aim in the second one (architecture vision) is to define and validate the principles, goals and strategies of the business. After having this business information, the next step is to determine the architecture's principles [16].

Based on [31], the framework focuses on three main activities: (i) Accessing knowledge (A), (ii) Obtaining knowledge (O) and (iii) Sharing knowledge (S). Those activities are matched against TOGAF's first and second phases, as shown in Tables 2 and 3, which shows the main activities proposed by TOGAF for the preliminary and vision phases along with the knowledge processes to be managed. Both Tables 2 and 3 use the letters A, O and S, as previously described.

Table 2. Activities in preliminary phase [6].

Define enterprise	Identify enterprise's elements	Define framework to use	Define tools and infraestructura	Define architecture principles
A	A			
O		O	O	
S		S	S	S

Table 3. Activities in architecture vision [6].

Define goals	Define architecture's scope	Define requirements	Define value proposal	Identify the impact
				A
O	O	O		
S		S	S	S

5.2 Way of Modeling

This framework includes activity modelling through the BPMN notation. This notation describes the way current processes are managed; this is needed in order to use it in the rest of the framework. The second step of this "Way of modelling" is to classify knowledge through the ontologies generated from the information of existing proposals. The modeling of knowledge through ontologies identifies the stakeholders who generate, access and use knowledge. It also identifies the knowledge that should be managed within the framework, the artifacts used and/or generated, and the relationship between them. In addition, ontologies are used to identify information that must be stored and displayed within the lessons learned system.

As this project is based on TOGAF's preliminary and architecture vision stages, we modelled the processes in those phases in order to identify those processes that are susceptible of management within the KMF. In this way, we could identify if there were changes on them that could affect current processes. Unfortunately, given the confidential nature of some of these processes, they cannot be explicitly reported.

The third step is taking into account the way in which knowledge is stored and/or made available. This is important because the documentation of every phase of ADM in TOGAF should be classified and the resulting knowledge must be available easier and faster for the rest of the process.

5.3 Way of Working

According to the activities identified in TOGAF-based proposals for EA projects, the "way of working" has five tasks described below. The first task is the classification of the architecture principles used in each project as well as the business requirements. The classification of the architecture principles identifies which of those principles can be reused. This classification also includes the type of business of the company for which the EA is done, the size and the scope of the project and if this project was planned or integrated with other EA frameworks. These characteristics are transformed into tags in the classification system. The other category, the business requirements, allows the reuse of existing solutions for similar requirement types.

The second task is designed for supporting the knowledge generation process, taking advantage of the results of the first task. Here, a classification of previous completed projects (with varying degrees of success), as well as proposals not carried out, is created. To do that, we take into account some factors (the tags of the classification) like financial success, development time, best practices, customer satisfaction, among others. This classification is supported by the first task. This process supports knowledge traceability, relating it with the projects in which it was generated.

The third task is oriented to reusing the knowledge. In this task, and taking the results of the first and second tasks as inputs, we identify the assigned roles for each project in order to know how they are intervening in each kind of project; in this way, we can have them on the "work table" for future projects.

Those tasks describe how the knowledge we have about projects under development or already developed is managed within the KMF, but this must be supported by a process of knowledge capture and storage, which allows obtaining knowledge in an orderly manner for subsequent optimal search. That's why we define the following two tasks.

The fourth task generates an orderly way to face the process of knowledge capture, focusing on the ADM activity in order to know the user company and to validate its mission, vision and goals. At this point, we use some of the information generating objects like emails, the user company web page and interviews of the members of the company in order to identify and classify in which project they were used.

The last task addresses the process of knowledge storage in order to facilitate its access, taking account the typical knowledge flow and maintaining its quality. Here we take advantage of the classification of the last task in order to keep the information generating objects (i.e. codified knowledge sources) according to their respective project.

5.4 Way of Controlling

According to [32], knowledge management performance may be placed in the context of the Balanced Scorecard. As such, they provide a method to measure the projects done for a company around the evaluation of four perspectives: financial, customer, internal processes and development and learning. This method controls a knowledge management process based on the intellectual capital to, in our case, align the strategy of the enterprise architecture area with the KMF which is been proposed.

This intellectual capital is monitored from the employee's perspective (financial perspective) of the project roles which manage knowledge and the roles which generate, make available and use knowledge in order to avoid the creation of personal dependencies when someone needs access to it. The customer perspective will be evaluated from the point of view of successful projects, especially according to codified factors and executed proposals. The internal processes perspective in this case includes the selected phases of TOGAF: the way in which these phases are currently done, the way in which we propose do them with the KMF and the technology (last perspective) will be faced from the point of view of the KMF's support in information technologies tools, which are presented in the next section. These perspectives may incorporate specific performance assessment tools, such as process mining for the internal perspective, customer satisfaction for customer perspective, financial performance for the financial perspective and acceptance and success models (e.g. TAM or DeLone and McLean) for the learning and development perspective.

5.5 Way of Supporting

The "way of supporting" of the KMF is about the instantiation of the framework in a KMS (Knowledge Management System) prototype. In the next sections, we describe the tools and resources used to design and build the prototype.

Definition of KMS

This definition is given by three phases: (i) existing search tools to support the knowledge management framework and the degree of encoding of the information within it; (ii) identification of tools or technologies to support the activities described in the way of working; (iii) definition of the system architecture to provide a reference model for the KMS implementation.

Existing Tools

According to the case study, we found tools currently used to support the initial phases of an enterprise architecture project based on TOGAF. These tools are presented in Table 4.

Table 4. Existing tools [6].

Tool	Use description
Search in company web page	In the early stages of TOGAF we seek to understand the company to which the project of enterprise architecture is implemented, for that reason, often this information initially is searched in the pages of customers. Besides that, we need to define the use of certain technologies of information for the support to the final architecture
Email systems	Through emails the company requests and provides information throughout the process in early stages
Repository	In the study case, currently the company has a system for documentation management supported on a SharePoint server, but in this certain documents some there are some documents without any specification or order. Additionally, many documents are also stored for each role in each of the computers they use In these repositories the company has a series of documents in which the learned lessons from each project are reflected, initially during the generation of the proposal and later in the project

New Tools or Technologies

In order to understand what kind of tools we needed in our framework and to comply with the requirements presented during the meetings with experts from the area EA of the company, we searched for tools in order to support the framework and its validation through the development of a software prototype. This search allowed us to find tools to classify information such as tagging, representation of information, lessons learned systems, and enterprise repositories. Among the tools that use tags we identified, as an important characteristic, the need for collaboration between users to share knowledge through keywords. Such collaboration allows an enterprise to have a classification evolve from emergent patterns, because some users will have more tags than others [33]. The use of tagging can be used by technologies such as “entity linking”, this is used in the framework called UnBWiki. UnBWiki identify in a text entities and words to get your relationship automatically [34].

Some of these tools can also be visual browsers such as Yasiv for Amazon [35], weave [36], Gephi [37], NodeXL [38], d3 data-driven documents [39]. These visual tools gave us some ideas to represent graphically the obtained information in the first and second tasks in the way of working of our framework. The goal is to show these classifications according to their labels and easy search.

Finally, we propose lessons learned systems in the development of the prototype. For instance, the Knoco System [40], shows that we can have services such as design, capture and learning obtained from the analysis of lessons.

This search helped us to decide that prototype should not affect the development time of current EA projects. For this reason, a system of lessons learned must be generated rapidly at the end of each project to provide feedback.

Development of a KMS Method

We propose to use ontologies as the principal tool to the development of KMS that supports the framework. These ontologies will be made from the combination of the ontologies development methods like "Methontology" – it's a method to generate ontologies from scratch – [41] and "On-To-Knowledge" – it's a methodology to build systems from ontologies – [42]. The methods have some activities that complement each other and give dynamic to the project because it need prototyping and refinement from expert analysis [43].

KMS Architecture

To define the architecture that will support the framework, we validated those that are applicable to the project because of the size and scope of the initial TOGAF phases and proposed by [44]. The architectures proposals were: (i) task-based, (ii) centralized and (iii) distributed (view Table 5).

Table 5. Architectures for KMS [6].

Architecture	Description
Task-based	This architecture allows modular KMS from the context, the articulation of tasks and processes, through a workflow, describing and classifying information sources, generating information acquisition, all supported by good management technologies information [44]. With the use of this architecture, we could take advantage of workflow management, context, and resources where the information comes. It will enable reuse these information for the organization
Centralized	This architecture can be exploited using a single server that allows access of information to all users of the project, taking account the high number of customers and of projects developed and will develop, [44]. This allows consolidate knowledge and is useful for knowledge management of small segments of the organization, requiring an infrastructure with high availability and processing especially when the information sources are very high
Distributed	It can generate KMS with direct communication between each of the members of the project from a peer-to-peer approach, where you can have instant messaging, document sharing and use of collaboration tools, but it must be supported by a systematic order, which allows reuse of knowledge

6 Validation of the Knowledge Management Framework for Enterprise Architecture

After the framework was designed, we made its validation with the aim to verify its behaviour in the process of building a real enterprise architecture. This validation was made based on [45]. The actual KMS prototype as a software tool will be reported elsewhere.

6.1 Construction and Validation of a Prototype with the Technology Acceptance Model - TAM [46]

In this phase, we developed a prototype to verify if the knowledge management of the lessons learned was possible in a process of EA construction. This prototype was implemented in the case study company and we applied a survey that was developed based on TAM. This survey had the aim of identifying the utility and ease of use of the framework, as perceived by the user. These perceptions helped to identify the attitude towards use and the use intention.

The prototype is a software that graphically represents the learned lessons obtained in the development of enterprise architecture projects. These lessons are stored and sorted by description, status, date of generation, phase in which it was presented and/or project. This storage is done inside the software through to enter, edit or delete the lesson's information. After having lessons, the software allows filtering them based on each item of classification. The result of this search is a graphic relationship between related lessons. For instance, two lessons generated in the same phase of EA will be linked. Those links generate a relationship graph of lessons where each kind of data has a different color: a link between two lessons generated in phase A will have a different color than a link of lessons generated in the preliminary phase. All this helps to understand what kind of lessons exist and how they are related inside an enterprise context.

6.2 Enterprise Architecture Experts' Opinion

During the development of the framework and its validation, two case study company experts provided information about real EA projects and gave us some requirements on knowledge processes. In addition, they performed a validation of both the framework and the prototype.

The experts were an engineer, enterprise architecture senior consultor on the study case company and a consulting area manager of enterprise architecture and member of the research committee of The OpenGroup - Latin America, who is in charge of the TOGAF's internalization.

The developed framework was presented to these two experts and they could use the prototype too. Their opinions were:

- The knowledge management of the prototype is really helpful to the company.
- They highlighted the visual facility and the use facility of the prototype in order to manage the lessons learned, especially because it generates a unified structure to identify best practices and minimizes errors. Also, this improves the availability and use of the lessons learned and identifies the lessons with relationships they have in common in order to define actions that must be repeated or mistakes that should not be repeated.
- Having a knowledge management process in order to access, obtain and share knowledge systematically enables feedback about the lessons learned in each of the projects.

- Once you have used the prototype, it is possible to determine that the use of a KM of completed or in progress projects are helpful for starting a new one.

In addition, the framework and its validation supports the management of knowledge gained in the development of the selected phases. Also, we found that once the prototype was used by experts, the requirements presented by them could be satisfied without significant additional workload or time.

7 Future Works and Limitations

The prototype was designed with two main limitations. First, the limitation to the two first phases in TOGAF, and, second, an expert validation based on potential (not actual) utility.

As future work, we propose the application of the framework in a full EA project. The next one is to obtain conclusions not only with expert's opinions, but also with measurement certain of indicators set for an EA project. Also, the framework can be extended to other phases of EA proposed by TOGAF, so that more knowledge processes can be identified and managed. Finally, we propose that the KM system use semantic web technologies to generate knowledge from documents, text, emails, among other artifacts developed in previous EA projects.

8 Conclusions

As it was identified during the conceptualization of the research, the knowledge management wants to increase the value of the services and products generated by the organization through the knowledge. For this, it is necessary that aspects of the organization as assets, people, processes, etc., transfer knowledge in an optimal way. It is also necessary to identify which processes, actions or lines of business are susceptible to the application of techniques and methodologies of KM. We identified that many times the KM is applied and/or used in process of the company, for example day-to-day processes or processes to determine how to align the use of ICT as the business architecture does it.

In addition, we found that the projects that generate an enterprise architecture for a client company did not have a KM to allow the knowledge to be reused according to best practices, learned lessons, organization of tasks, etc.

In the research project we define which processes are likely to be managed. The creation of knowledge in consulting projects in business architecture was identified as the first process to be managed, so the EA area has the opportunity to reuse knowledge. In addition, it was also identified which processes already carried out knowledge management and brought the organization advantage afterwards or which processes needed a change.

The development of this project with an important IT company with an exclusive TOGAF EA area, gave us the opportunity of improving knowledge management in a real case of study. Thanks to that, we designed, built and validated a solid knowledge management framework.

This framework supports the generation, use, reuse and store of knowledge, in this case for the initial TOGAF phases, this allow company to make decisions in subsequent proposals or development of new projects. This supports a cycle of knowledge management that allows to company to know its processes, the way how they are developed and the way how the company can implement the framework in order to do the processes better.

The framework proposed improves the management of the information therefore gives knowledge to the company that is used as experience for new projects, because this framework organizes the identified knowledge in the initial phases of the project through tagging and ontology engineering. This framework generates, reuses, adds and stores knowledge, because it is aligned with the company's current processes that are in turn defined by a standard framework like TOGAF. The proposed framework doesn't generate extra work for the company because they are embedded in the "normal" processes; for that reason, we propose the use of the framework in order to support the knowledge management in EA projects in an orderly and productive way.

The ability to manage explicit knowledge, generated for the selected phases of a business architecture project based on TOGAF, was identified. This explicit knowledge is reflected in the proposals and the lessons learned generated in the development of the project.

We learned with the investigation the implementation of KM must count with a well-defined limits in order to get an integral management of the processes, for instance we selected two phases of TOGAF, because working with whole organization or all its projects can bring problems about the specifications for an KM because of the amount of data, information, knowledge and/or resources. As to the obtained results, the KM framework validation with a software prototype also enabled potential users, who are not experts in KM, to understand how knowledge can be managed in a tangible way, getting visible results for the organization's EA area.

Lastly and according to experts' comments, it is possible to apply the same management for the rest of the phases of a project developed under TOGAF, since this also generates lessons learned.

References

1. Lankhorst, M.: Enterprise Architecture at Work: Modelling, Communication and Analysis. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-29651-2
2. Arango Serna, M.D., Londoño Salazar, J.E., Zapata Cortés, J.A.: Arquitectura empresarial - una visión general Rev. Ing. Univ. Medel. **9**(16), 101–111 (2011)
3. Struck, V., Buckl, S., Matthes, F., Schweda, C.M.: Enterprise architecture management from a knowledge management perspective-results from an empirical study. In: MCIS, p. 84 (2010)
4. The Open Group: TOGAF Version 9. The Open Group (2009)
5. Ernst, A.M.: Enterprise architecture management patterns. In: Proceedings of the 15th Conference on Pattern Languages of Programs, New York, NY, USA, pp. 7:1–7:20 (2008)

6. Meneses-Ortegon, J.P., Gonzalez, R.A.: Knowledge management framework for early phases in TOGAF-based enterprise architecture. In: Proceedings of 8th International Joint Conference on Knowledge, Discovery, Knowledge, Engineering and Knowledge Management, KMIS, vol. 3, pp. 31–40 (2016)
7. Filemon, A., Uriarte, J.R.: Introduction to Knowledge Management, Supported by National Academy of Science and Technoogy. ASEAN Foundation (2008)
8. Yew Wong, K., Aspinwall, E.: An empirical study of the important factors for knowledge-management adoption in the SME sector. J. Knowl. Manag. **9**(3), 64–82 (2005)
9. Alavi, M., Leidner, D.E.: Review: knowledge management and knowledge management systems: conceptual foundations and research issues. MIS Q. **25**(1), 107 (2001)
10. Baskerville, R., Dulipovici, A.: The theoretical foundations of knowledge management. Knowl. Manag. Res. Pract. **4**(2), 83–105 (2006)
11. Pandey, K.N.: Paradigms of Knowledge Management: With Systems Modelling Case Studies. Springer, Berlin (2016). https://doi.org/10.1007/978-81-322-2785-4
12. Dalkir, K.: Knowledge Management in Theory and Practice. Routledge, London (2013)
13. Allameh, S.M., Zare, S.M., Mohammad Reza Davoodi, S.: Examining the impact of KM enablers on knowledge management processes. Procedia Comput. Sci. **3**, 1211–1223 (2011)
14. Rubenstein-Montano, B., Liebowitz, J., Buchwalter, J., McCaw, D., Newman, B., Rebeck, K.: A systems thinking framework for knowledge management. Decis. Support Syst. **31**(1), 5–16 (2001)
15. Wegmann, P.A.: On the systemic enterprise architecture methodology (SEAM, in SEAM). Published at the International Conference on Enterprise Information Systems 2003 (ICEIS 2003), pp. 483–490 (2003)
16. The Open Group: TOGAF® Version 9.1, an Open Group Standard (2011). http://pubs.opengroup.org/architecture/togaf9-doc/arch/
17. Mezzanotte, D.M.: Planning enterprise architecture: creating organizational knowledge using the theory of structuration to build information technology. 2016 IEEE ACIS 14th International Conference on Software Engineering Research, Management and Applications (SERA), pp. 107–115 (2016)
18. Mihi-Ramirez, A., Garcia Morales, V.J., Marti Rojas, R.: Knowledge creation, organizational learning and their effects on organizational performance. Inzinerine Ekon.-Eng. Econ. **22**(3), 309–318 (2011)
19. Moscoso-Zea, O., Lujan-Mora, S., Esquetini Caceres, C., Schweimanns, N.: Knowledge management framework using enterprise architecture and business intelligence. In: Proceedings of International Conference on Enterprise Information Systems, ICEIS, pp. 244–249 (2016)
20. Buckl, S., Matthes, F., Schweda, C.M.: Future research topics in enterprise architecture management – a knowledge management perspective. In: Dan, A., Gittler, F., Toumani, F. (eds.) ICSOC/ServiceWave - 2009. LNCS, vol. 6275, pp. 1–11. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16132-2_1
21. Del Nostro, P., Orciuoli, F., Paolozzi, S., Ritrovato, P., Toti, D.: ARISTOTELE: A Semantic-Driven Platform for Enterprise Management. IEEE, New York (2013)
22. Moscoso-Zea, O., Andres-Sampedro, Lujan-Mora, S.: Datawarehouse design for educational data mining. In: 2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET). IEEE, New York (2016)
23. Migdadi, M.: Knowledge management enablers and outcomes in the small-and-medium sized enterprises. Ind. Manag. Data Syst. **109**(6), 840–858 (2009)
24. Tu, F., Yang, C., Xiao, H., Yang, D.: A study on typical architecture layers of the digital library alliance based on the enterprise architecture, pp. 142–145 (2012)

25. Fiedler, M., Hauder, M., Schneider, A.W.: Foundations for the integration of enterprise wikis and specialized tools for enterprise architecture management. Presented at the Wirtschaftsinformatik, p. 109 (2013)
26. Anajafi, F., Nassiri, R., Shabgahi, G.L.: Developing effective project management for enterprise architecture projects. In: 2010 2nd International Conference on Software Technology and Engineering (ICSTE), vol. 1, pp. V1-388–V1-393 (2010)
27. Buckl, S., Matthes, F., Schweda, C.M.: Future research topics in enterprise architecture management - a knowledge management perspective. In: Dan, A., Gittler, F., Toumani, F. (eds.) ICSOC/ServiceWave - 2009. LNCS, vol. 6275, pp. 1–11. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16132-2_1
28. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A design science research methodology for information systems research. J. Manag. Inf. Syst. **24**(3), 45–77 (2007)
29. Indra. http://www.indracompany.com/. Accessed 03 Dec 2014
30. Opt'Land, M., Proper, E., Waage, M., Cloo, J., Steghuis, C.: The results of enterprise architecting. In: Opt'Land, M., Proper, E., Waage, M., Cloo, J., Steghuis, C. (eds.) Enterprise Architecture, pp. 49–83. Springer, Berlin (2009). https://doi.org/10.1007/978-3-540-85232-2_4
31. Rus, I., Lindvall, M.: Guest editors' introduction: knowledge management in software engineering. IEEE Softw. **19**(3), 26–38 (2002)
32. Fairchild, A.M.: Knowledge management metrics via a balanced scorecard methodology. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences, 2002, HICSS, pp. 3173–3180 (2002)
33. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32**(2), 198–208 (2006)
34. Monteiro, L.B., Weigang, L., Saleh, A.A.: An approach of vector space model to link concrete concepts with Wiki entities. Presented at the Proceedings—15th IEEE International Conference on Computer and Information Technology, CIT 2015, 14th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2015, 13th IEEE International Conference on Dependable, Autonomic and Secure Computing, DASC 2015 and 13th IEEE International Conference on Pervasive Intelligence and Computing, PICom 2015, pp. 313–320 (2015)
35. Amazon: Amazon Products Visualization - YASIV (2012). http://www.yasiv.com/. Accessed 24 May 2016
36. University of Massachusetts Lowell: Weave (Web-based Analysis and Visualization Environment) (2015). http://www.oicweave.org/. Accessed: 14 June 2016
37. Gephi: The Open Graph Viz Platform (2010). https://gephi.org/. Accessed 14 June 2016
38. Microsoft: NodeXL® Network Overview, Discovery and Exploration for Excel, CodePlex (2015). http://nodexl.codeplex.com/Wikipage?ProjectName=nodexl. Accessed 14 June 2016
39. Bostock, M.: D3.js - Data-Driven Documents (2015). https://d3js.org/. Accessed 14 June 2016
40. Knoco: Lessons Learned Services—Knoco Ltd. (2014). http://www.knoco.com/lessons-learned-introduction.htm. Accessed 14 June 2016
41. Corcho, O., Fernández-López, M., Gómez-Pérez, A., López-Cima, A.: Building legal ontologies with METHONTOLOGY and WebODE. In: Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A. (eds.) Law and the Semantic Web. LNCS (LNAI), vol. 3369, pp. 142–157. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-32253-5_9
42. Corcho, O., Fernández-López, M., Gómez-Pérez, A.: Ontological engineering: principles, methods, tools and languages. In: Ontologies for Software Engineering and Software Technology, pp. 1–48 (2006)

43. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web (2004)
44. Maier, R.: Knowledge Management Systems: Information and Communication Technologies for Knowledge Management. Springer, Berlin (2007). https://doi.org/10.1007/978-3-540-71408-8
45. González R.A., Sol, H.G.: Validation and design science research in information systems, pp. 403–426 (2012)
46. Varela, L.A.Y., Tovar, L.A.R., Chaparro, J.: Modelo de aceptación tecnológica (TAM): un estudio de la inf luencia de la cultura nacional y del perfil del usuario en el uso de las TIC, 2010, pp. 187–204 (2010)

Empowering SMEs to Make Better Decisions with Business Intelligence: A Case Study

Raghavendra Raj[1(✉)], Shun Ha Sylvia Wong[2], and Anthony J. Beaumont[2]

[1] AGGORA Group, North Moons Moat, Redditch B98 7UX, UK
raghavendra.raj@aggora.com
[2] School of Engineering and Applied Science, Aston University, Aston Triangle, Birmingham B4 7ET, UK
{s.h.s.wong,a.j.beaumont}@aston.ac.uk

Abstract. With the advance of Business Information Systems (BIS), irrespective of the size, companies have adopted an approach to electronic data collection and management for two decades. The advancement in technology means they have in their possessions large volumes of historical data. Large organizations have cached on this and use a range of tools and techniques to leverage the usefulness of this information to make more informed business decisions. For most small and medium-sized enterprises (SMEs), however, such data typically sits in an archive without being utilized. While SMEs appreciate the need for utilizing historical data to make more informed business decisions, they often lack the technical knowhow and funding to embrace an effective BI solution. In this paper, drawing from our experience in implementing a BI solution for a UK SME we discuss some potential tools and strategies that could help SMEs overcome these challenges so as to reap the benefits of adopting an effective BI solution.

Keywords: Business Intelligence · Data warehouse · Microsoft BI
SME

1 Introduction

Information Technology (IT) has become an essential for businesses of any size for over three decades. For most businesses, this has facilitated the collection of a vast amount of business transaction data. While such data are important to support smooth operations of the company at the time of its creation, once the respective business transactions have been dealt with, they are often being archived away and are unlikely to ever be revisited. However, such historic data, when analyzed appropriately, can provide important clues to discover new business opportunities and to improve the company's business processes. The ability to efficiently manage, access and analyze large volumes of company historic data means that business decision making is more informed and business trends and risks can be more easily identified.

Business Intelligence (BI) rediscovers the usefulness of existing business information. It equips managers and decision makers with important information to perform

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 306–325, 2019.
https://doi.org/10.1007/978-3-319-99701-8_15

business analyses that are needed for making key business decisions. In large corporations, BI has been one of their core strategies for growth for more than twenty years. With the rapid changes in business climate and conditions, even small and medium-sized enterprises (SMEs) have increasingly look to adopting BI in supporting their business decision making process. However, when looking to adopting a BI solution, many SMEs encounter issues such as a general lack of technical expertise to convert transaction data into business information and the general lack of funding to invest in a suitable BI solution. Furthermore, the lack of understanding of the benefits of BI also makes SMEs reluctant to invest in the adoption of new BI solutions. Without the support of quality data, the decision-making in SMEs resorts to relying on the results from various Information and Communications Technology (ICT) tools built as a part of the company's infrastructure, but were not designed to perform business analysis. This means that the results may not be accurate and led to suboptimal business decisions being made.

Business Intelligence is not a novel concept. This term was first put forward by Luhn [15] and was reintroduced by Howard Dresner [2]. As Negash and Gray [24] explained, BI systems are specialist IT systems which "*combine data gathering, data storage, and knowledge management with analytical tools to present complex and competitive information to planners and decision-makers*". Such systems typically analyze data from a centralized data repository which hosts business and company data aggregated from various sources. Due to the advancement of BI tools, the adoption of BI has grown significantly since then. Furthermore, the readiness of companies to adopt new strategies to stay ahead of their game also pushes up the demand for high quality business intelligence. This has led to more concerted efforts being placed in developing new BI technologies. Recent advances in technologies such as Decision Support Systems (DSS), Executive Information Systems (EIS), Data Warehouse (DW), Online Analytical Processing (OLAP) and Data Mining have also improved the capabilities of BI systems and have played a significant role in raising awareness, and also in increasing popularity, of Business Intelligence in the business sector.

With the increased popularity of BI amongst businesses, proven benefits of adopting BI in businesses have emerged. For example, Watson and Wixom [29] reported that implementing BI solutions could lead to faster and cheaper information retrieval, thus helping organisations to achieve their business goals. Howson [11] noted that BI helped employees in all divisions to interact with, and analyze, business data in order to facilitate a more informed business management process. This has led to an increase in company efficiency, the ability to identify new business opportunities and the ability to facilitate operation reengineering. A survey of more than 2,300 CIOs published by Gartner in 2012 [6] revealed that BI was ranked the top technology priority by the participating companies as BI enabled them to create new capabilities. As summarized by Chugh and Grandhi [3], the key benefits of implementing BI solutions in an enterprise includes:

- Equipping the company with the ability to analyze data from multiple sources and using different dimensions;
- Enabling managers to make informed business decisions through identifying important patterns of behaviour captured in the data;

- Improving accuracy in predictions;
- Helping the company to identify root causes of problems so as to improve operational efficiency.

A good range of businesses has been utilizing BI to assist in making key business decisions for many years, e.g. banking, financial services, health service, IT companies, insurance companies, manufacturing industry, etc. A majority of large organizations are already drawing benefits from using BI. For many of them, BI has become one of the major strategies to maintain a competitive edge. Furthermore, contrary to what it may seem, BI is not just for top-level management such as Managing Directors. When appropriately implemented, BI can empower a wide range of business decision-making processes. As was reported by Horakova and Skalska [10], BI is not restricted to top-level management, different company departments and business units use some kind of BI tools and the main users of BI include managers from sales, marketing, purchasing, accrual, finance, accounting, human resource and IT.

1.1 BI Solutions

BI solutions comprise multiple technological components. To generate business intelligence, data from business transactions and other relevant company processes needs to undergo an Extract-Transform-Load (ETL) process. This process cleanses, transforms and restructures business data and stores it in a data warehouse or a data mart. Data warehouses and data marts are essentially databases specially designed for promoting data analysis and knowledge discovery. How to analyse the data is dependent on the needs of the company and hence varies from business to business. One common area of data analyses is driven by the company's Key Performance Indicators (KPIs). To assist in knowledge discovery, data visualization tools are often used to present the results of data analyses to BI users. These results may be presented as standard KPI reports generated from regular queries or as on-demand reports that are generated for informing specific business decisions (Fig. 1).

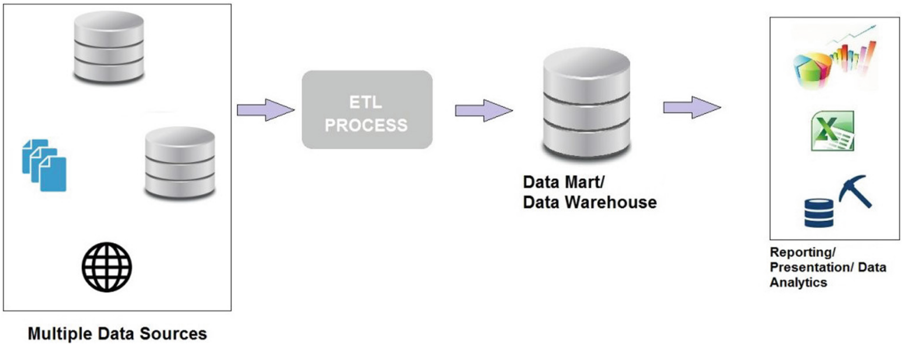

ETL PROCESS
Data Mart/ Data Warehouse
Reporting/ Presentation/ Data Analytics
Multiple Data Sources

Fig. 1. Key BI components and processes [26].

In the past, each BI component was developed and supplied by specialist vendors. Selecting the right combination of BI components that would meet the needs of a business requires technical expertise that is often not readily available in the company. The need to integrate various BI components also pushes up the overall implementation cost of a BI solution. In the last decade, with a paradigm shift in the way software components communicate with each other, suppliers have developed, a new generation of BI components have been developed with interoperability in mind and hence could be easily integrated. The cost of implementing BI has therefore been reduced significantly. Some of the major suppliers of BI solutions are well-known names such as Oracle, Microsoft, IBM, SAS and SAP. More recently, BI solutions from smaller specialist vendors such as Qlik and Tableau are gaining traction. To make BI solutions more accessible, some major suppliers of technological solutions such as Microsoft and IBM have begun to supply BI products which cover the entire stack of BI components. Some of these suppliers even deliver BI components as part their standard business solution with no extra cost. For example, a standard suite of Microsoft BI components is included as part of the Microsoft Developer Network (MSDN) license [21].

BI solutions do not come in locally-hosted, server-based form only. Some contemporary BI solutions take the form of Software as a Service (SaaS), which is essentially a cloud-based BI solution. As there is no need for sourcing software and hardware for implementing and hosting the BI solution, the implementation and maintenance costs for SaaS-based BI solutions are minimal. With SaaS, companies simply pay monthly or yearly subscriptions to access the cloud-based service over the Internet. Company data is typically stored in the cloud and managed by the service provider [25]. As new versions of the cloud-based service becoming available, companies using SaaS will be able to utilise the new features instantly. One typical application of SaaS-based BI solutions is in the area of Customer Relationship Management (CRM) where the generated intelligence on customer satisfaction is used to improve customer services.

With the availability of inexpensive BI options and a wide range of BI solutions for business use, when adopting BI, what a company needs to do is to identify a suitable strategy to integrate an appropriate BI solution (or components) within the company's existing IT infrastructure. This requires certain level of technical know-how that may be more readily available in large organisations than small and medium-sized enterprises (SMEs).

2 BI in SMEs

There is a general opinion that SMEs are trailing behind in adopting Business Intelligence to assist in their decision-making. According to a survey conducted by McCabe [16], 33% of medium-sized companies adopted some kind of BI solutions, and a further 28% of them planned to take advantage of a BI solution. Amongst smaller organizations, however, only 16% adopted a BI solution and a further 16% planned to use a BI solution. While McCabe's survey showed that the adoption of BI amongst SMEs was slow, it also reported that there was an upward trend in the awareness of the need for BI.

The European Commission (EC) defined a small and medium-sized enterprise (SME) as a business with <250 staff and a turnover of $\leq$€50m. About 99% of businesses in the EU are SMEs [4]. With a relatively small turnover, SMEs typically do not have additional financial and human resources to invest in new, non-business-critical, technologies such BI systems. It therefore comes as no surprise that the adoption of BI amongst SMEs has been slow. Given the cost and complexity involved in adopting a specialist BI solution, many SMEs simply integrate their database with a spreadsheet software such as Microsoft Excel to produce some form of business intelligence [16, 28]. However, as standard spreadsheet software is not equipped with specialist features nor visualization tools to support data analysis and knowledge discovery, the resulting level of intelligence produced is rudimentary and is inadequate to provide a clear view of the company's current performance. Furthermore, for many SMEs, the term Business Intelligence is often misperceived as a technology for large organizations only. This led to a lack of incentive to explore the potential applicability of BI in their businesses. As a result, most decisions made by top level management in SMEs are based on information obtained from various ICT tools built as part of the company's infrastructure [28].

Irrespective of size and sector, there is a general appreciation of the necessity for, and the benefits of, using business intelligence to improve the company's business decision-making process. With the complexity involved in typical BI implementation processes and the lack of appropriate resources, the risk of failure amongst SMEs is high. Hence, many SMEs are put off adopting specialist BI solutions. Moreover, while there are numerous studies reporting successful adoption and utilization of BI amongst large organizations, reports on similar successful stories amongst SMEs are uncommon [5]. To promote the uptake of BI solutions amongst SMEs, more needs to be done to raise the awareness of the benefits BI can bring to SMEs and more effort needs to be placed on overcoming the initial challenges faced by SMEs when adopting specialist BI solutions.

2.1 Benefits of Adopting BI in SMEs

SMEs often operate in a competitive marketplace and under a relatively tight profit margin. To give the business a competitive edge, the management needs to keep abreast of a variety of key business information such as market trend, company performance and its clients' needs. Such information plays an important role in ensuring appropriate company strategies are developed and sound business decisions are made. Specialist BI solutions are designed to generate those kinds of important business intelligence from existing business data.

According to a research by Scholz et al. [27], SMEs can benefit from utilizing BI tools in many ways and the most important ones being: (i) improvements in data support, (ii) improvements in decision support, and (iii) cost and time saving.

In a BI solution, as a data warehouse is designed to facilitate data analysis and retrieval, easy access to business data is therefore guaranteed. The ETL process also ensures that business transaction data is cleansed and validated before entering the data warehouse. Such improvements in data support lead to improved data quality and help ensure the correctness of generated reports.

The visualization and data analytic tools in a BI solution provide rich visuals to help the management better-understand existing business data. Such an improved understanding of the business promotes an accelerated decision-making process. Furthermore, BI tools also help identify risks and hence leading them to be rectified in a timing manner.

One typical feature of BI tools is to generate visual output of business data analysis that are easy to interpret, e.g. in form of dashboards and scorecards. This provides managers and decision-makers with a quick way to identify potential issues within the business.

2.2 Challenges of BI

When considering whether or not to adopt a BI solution, SMEs are often confronted with the following issues:

- BI solutions are often expensive. For example, a cloud-based BI solution typically costs at least USD$500 per month per user. Even with a management team of moderate size, such a monthly cost would add a significant financial burden to the business.
- While off-the-shelf BI tools are available, for non-technical business users, the learning curve of such tools is often too steep to be achievable.
- Hosting a BI solution requires the support of a non-trivial, and often costly, hardware infrastructure.
- While there is a wide range of BI solutions available, SMEs often lack in-depth knowledge of BI to select an appropriate solution for meeting the business's needs.
- Generating BI is often a non-trivial task. It requires advanced knowledge and good understanding of database modelling and data warehousing. Such technical knowledge is often not readily available within most SMEs.

In summary, the challenges can be narrowed down to two main factors: lack of budget and insufficient technical know-how. While these factors pose significant challenges to SMEs in adopting a BI solution, they need not become the barriers to adoption of BI solutions amongst SMEs. To overcome the budget issue, for example, SMEs need to identify low cost BI solutions that will meet their needs. To this end, SMEs may consider adopting an IT solution that comes with a standard BI solution at no extra cost, e.g. Microsoft Office 365 [22]. However, many SMEs may not be aware of such a low-cost BI opportunity nor have the expertise required to start utilizing it. Though to achieve a richer range of data visualization, there may be the need to augment a standard BI solution with data visualization tools. However, the cost of purchasing data visualization tools can be reduced when lightweight web technologies are used to present the results. This, not only minimises the overall cost, it also removes the need for extensive training.

Reflecting on our successful experience with implementing a BI solution for an SME in the UK, in the following sections, we present our adopted approach and discuss how SMEs may overcome the challenges in implementing an effective BI solution.

3 BI Implementation: A Case Study

In a competitive business environment, SMEs are increasingly looking for new ways to improve their business decision making process. AGGORA [1], being an SME specializing in providing catering equipment solutions for the food service industry, teamed up with a UK university to create a novel IT platform that aims to deliver new functionality to support their business growth through the use of improved business intelligence. This case study reports the approach adopted in this project.

3.1 About the Company

UK based company AGGORA specializes in sales and service of catering equipment. Their clients range from major corporates to small businesses. The company has about 170 staff and has a turnover of approximately £25m a year. The market place for AGGORA has become more competitive lately and hence the company recognized the need to bring a paradigm shift in their business model, adding IT and servitization expertise to its traditional core strength in equipment and fittings.

AGGORA has grown significantly in recent years and the company plans to continue their rapid growth. To achieve this, the company planned to offer innovative services, extending its range of offers far beyond the traditional design and fabrication, equipment sales, and service and maintenance approach of competitors. They also wanted to have a flexible data management system that will provide valuable performance information on kitchen equipment and fittings from data collected by the company's flagship Asset Management System. Though the in-house IT systems have the capabilities to generate appropriate information for business analyses and planning, the report generation process was not sufficiently efficient to support the business needs.

AGGORA's leadership team understood the need for radically transforming the business to meet the rapidly changing needs of their marketplace and hence they looked to implement a BI solution to:

(a) support their expansion plans,
(b) improve their decision making process, and
(c) improve the quality of reports.

The main obstacle which confronted AGGORA was that the company did not possess sufficient technical expertise in BI technologies among its human resource pool and also retraining existing staff was not an option due to time and budget constraints.

To overcome these challenges, AGGORA joined a UK government's scheme called the Knowledge Transfer Partnerships (KTP) scheme. KTP is a UK-wide programme that has been helping businesses for the past 40 years to innovate and grow through access to funding as well as knowledge, technology and skills within UK academic institutions. Each KTP project is a three-way partnership between an SME, an academic institution and a recent graduate. The core aim of all KTP projects is to investigate and implement effective solutions for the identified business challenges. The goal of such a scheme is to transfer research knowledge into SMEs so as to bring about business growth [12]. Every KTP project includes a detailed plan for embedding the knowledge and skills into the company through workshops, training and detailed documentations. This helps ensure maintainability and sustainability of the project on completion.

AGGORA secured the funding for a KTP project and this enabled a graduate to be employed and work on implementing a BI solution with guidance from the partner university. As per the KTP norm, the graduate was employed by the university, but was based in the company so as to maximize involvements from all project stakeholders throughout the project development, ensuring that all views and concerns considered and addressed, from the beginning of the project. This helped ensure the design of a BI solution that would address the business needs and a smooth roll out of the BI solution.

3.2 Project Objectives

The aims of the KTP project are:

(a) to increase the scope of data available for reporting,
(b) to minimize the involvement of the IT team by empowering managers to create and maintain their own reports,
(c) to create a user-friendly reporting environment, and
(d) to lay foundation for a more sophisticated BI solution.

The goal of the project is to implement a suitable BI solution that will empower decision-makers within AGGORA to make more informed business decisions. Figure 2 summarized the goals of the intended BI solution.

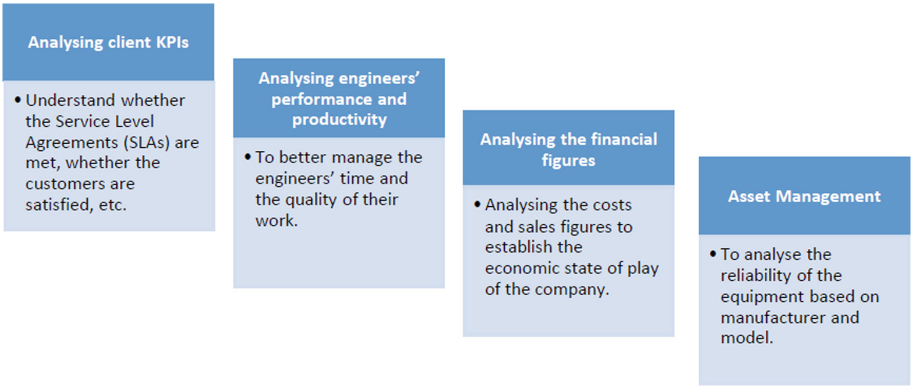

Analysing client KPIs
• Understand whether the Service Level Agreements (SLAs) are met, whether the customers are satisfied, etc.
Analysing engineers' performance and productivity
• To better manage the engineers' time and the quality of their work.
Analysing the financial figures
• Analysing the costs and sales figures to establish the economic state of play of the company.
Asset Management
• To analyse the reliability of the equipment based on manufacturer and model.

Fig. 2. BI project goals.

3.3 Implementation Approach

The implementation approach for a BI solution could significantly vary depending on multiple factors such as the size of the company, budget available for the project and the company's knowledge in implementing a BI solution. Larger organizations with sufficient budget may prefer to buy an off-the-shelf BI solution from a leading vendor.

Such solutions would typically offer staff training and support. However, such a solution might be too general and include features that are irrelevant to the organizations. On the other hand, an organization with large IT budget may choose to employ a team of experts to implement and maintain a bespoke BI solution. However, such a solution is unlikely to be affordable and manageable by an SME such as AGGORA.

AGGORA recognizes the need to utilise a BI solution for providing insight into the data from all aspects of its business. However, the company also recognized that implementing a suitable BI solution may not be straight-forward and might have mixed outcomes, including long delays, budget overruns, data problems, and dissatisfied end users.

While trying to accomplish all the requirements and deliver the entire solution all at once may sound ideal, it is unlikely to work well for an SME f or the following reasons:

- Changes in business rules and desired metrics may cause delays and conflicts in development.
- There may be the need to resolve legacy data validation issues before the core development commences.
- Requirements could be misinterpreted and hence the deliverables do not meet the requirements.
- There may be significant delays in the development due to changes in business environment and company priorities.

Prompted by constant changes in business priorities and the need for flexibility, we opted for a five staged iterative approach suggested by McGonagle [17] in our BI solution development. Figure 2 shows our adopted approach. The project lasted for 27 months and the BI solution development went through three iterations (Fig. 3).

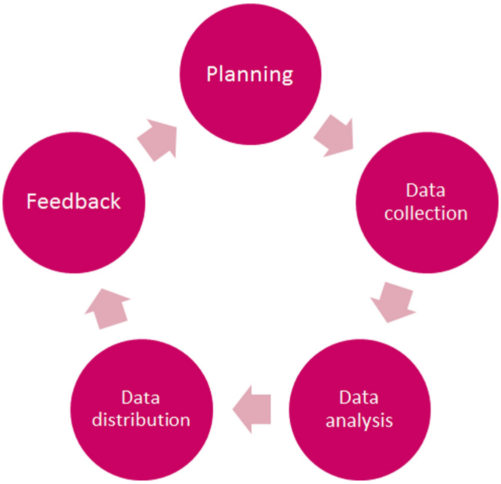


Fig. 3. BI implementation approach adopted by AGGORA.

Planning. The implementation started with the planning phase. It is a good practice to involve all stakeholders from the management team from the get-go and gather the analysis and reporting requirements up front. In this phase, the management team who would be consuming the intelligence were invited to articulate their requirements and establish the Key Performance Indicators (KPIs) required to be measured. Having all the necessary requirements from the users up front saves a lot of extra rework in the later stages. This also helps to scope out the project and the collected requirements also form the foundation for the subsequent phases.

As there are different tools available for implementing a BI solution, we decided to use the Microsoft suite of BI tools. This decision was prompted by the existing IT infrastructure available in the company. Since the company already uses Microsoft SQL Server 2012 as its backend database along with other Microsoft business products, the Microsoft BI suite facilitated a seamless integration with the company's IT infrastructure. The various components on a Microsoft SQL Server 2012 [20] supported our implementation of the BI solution are:

- **SQL Server Database Engine** – This includes the Database Engine, the core service for storing, processing and securing data.
- **Integration Services** - This is a set of graphical tools and programmable objects for moving, copying and transforming data.
- **Analysis Services** - This includes the tools for creating and managing online analytical processing (OLAP) and data mining applications.
- **Reporting Services** – This includes server and client components for creating, managing and deploying reports.
- **Master Data Services** – This is the SQL Server solution for master data management.

Data Collection. Once the requirements are gathered and objectives are set, we need to transform the data into a format that can be consumed by analytic applications. It was decided that the data from the source would be transformed into and stored in a data warehouse. A data warehouse is fundamentally a database, although there are some significant differences between the design process and best practices for an online transaction processing (OLTP) database and a data warehouse that will support online analytical processing (OLAP) and reporting workloads. With traditional business information systems, the focus is to process business transaction data. Such data is typically stored in a relational database that is designed for facilitating OLTP and is optimized for data entry, retrieval and general transactional processing. With an information system that provides business intelligence, the focus is knowledge discovery and reporting, with online analytic processing being its main task. While these two types of information systems are designed to work with the same data source, their design approaches are very often different due to the need to fulfil different requirements.

One core task in the data collection phase is to identify the business metrics that will drive the data analysis. These metrics are known as the dimensions of the analysis, and the procedure is called dimensional modelling which is a very popular data analytic technique used in the design of data warehouses. The main reasons for its popularity are that it brings about fast query performance and it also presents data in a user consumable format. Although data warehouses can be implemented as normalized relational database schemas, most designs are based on the dimensional model advocated by Kimball [13]. In a dimensional model, the numeric business measures are stored in fact tables. Each fact table is typically linked to multiple dimension tables that contain the attributes by which the measures can be aggregated.

We have followed Kimball's four steps to dimensional modelling [13]:

- **Select the Business Process.** Business processes are the various operations performed by an organization. Every business process generates events that can be translated into fact tables. Choosing the set of business processes that would be most beneficial for business decision process is paramount. In this project, we have chosen the set of business processes that are relevant to the KPIs we identified in the planning stage.
- **Declare the Grain.** Declaring the grain is the pivotal step in a dimensional design. The grain defines exactly what should be in a single row of a fact table. The grain should be declared before identifying facts and dimensions to ensure consistency in the design. This consistency is critical for ensuring high performance and ease of use of the resulting BI solution. As a rule of thumb, defining the finest level of grains enables a wider range of business intelligence to be generated. Businesses should choose a level of granularity suitable for their requirements.
- **Identify the Dimensions.** Dimensions specify the "who, what, where, when, why, and how" of business process events. These are essentially the attributes used for filtering and grouping the facts. They contain the descriptive labels that enable the information from a Data Warehouse/Business Intelligence system to be consumed for business analysis. In this project, key parameters such as the customers and the assets for which the performance needs to be measured against were defined as the dimensions.
- **Identify the Facts.** Facts are the numeric values generated from a business process event. They contain all the measurements needed to provide answers to business questions. Every row in a fact table should be consistent with its corresponding grain. The facts we identified for this project were the numeric measures generated by events in the business process, such as an engineer visiting a client in response to a request to repair equipment.

As proposed by Kimball [13], a data warehouse could be based on two kinds of schemata:

- *Star Schema* usually consists of fact tables linked to dimension tables using primary/foreign key relationships.
- *Snowflake Schema* consists of hierarchical relationships in a dimension table, with normalised, low-cardinality attributes appearing as secondary tables connected to the base dimension table by an attribute key.

For the purpose of this project, we have opted for a hybrid approach which is a combination of both star and snowflake schema, partly because we had to integrate with an existing transaction database that was not designed with data warehousing in mind. We have used an iterative approach to build the data warehouse, and in some of the iterations a star schema was more appropriate and in others a snowflake schema was more appropriate. In following such a hybrid approach, we needed to slightly increase the complexity of the resulting data warehouse in order to maintain the data consistency among the shared dimensions. The trade-off for that is a slightly degraded performance although, due to the nature of our data, that slight degradation in performance was estimated to be less than 10% and not noticeable by end users.

Data Analysis. Once the data warehouse is designed, it has to be populated with data from the live transaction database. This is known as an Extract-Transform-Load (ETL) process.

There are several ways to implement an ETL solution, but SQL Server Integration Services (SSIS) is the primary ETL tool for SQL Server. SSIS includes:

- **SSIS Designer.** A graphical design interface for developing SSIS solutions in the Microsoft Visual Studio development environment.
- **Wizards.** Graphical utilities which enable developers to quickly create, configure, and deploy SSIS solutions.
- **Command-line Tools.** Utilities for managing and executing SSIS packages.

In our project, the source business data is generated from an in-house field service management system and the data is stored in the form of a relational database using Microsoft SQL server. We started by familiarizing ourselves with data source and designed the dimensional modelling for our data warehouse as described above. In order to populate the data warehouse, we implemented an ETL process by using SSIS. This process consisted of three SSIS packages, with each containing one or more Data Flow tasks. The key steps involved in an ETL process are:

a. ***Extract***
   To extract data, an SSIS package must be able to connect to the data source. In an SSIS solution, we defined data connections by creating a connection manager for each data source. As discussed earlier, our source data is the in-house IT system implemented as a single SQL Server database.
b. ***Transform***
   Data transformations enabled us to perform operations on rows of data as they pass through the pipeline. The transformations performed in our ETL process include:

- **Row Transformation**
  This task deals with the copying of business transaction data to the data warehouse and it is supported by three standard SSIS functions: copy, data conversion and derived column. The *Copy Column* function adds new columns which are the copies of columns from input data set. The *Data Conversion* function enables changes to the data type of a column to be made during the translation. The *Derived Column* function creates new columns derived from the values in the input columns.

- **Rowset Transformation**
  This task deals with aggregation of atomic data for facilitating data analysis and reporting. The SSIS package provides an *Aggregate* function which applies aggregation (minimum, maximum, average, sum, etc.) on the incoming set of data.
- **Split and Join Transformation**
  To support data analysis and knowledge discovery, various ways of dividing and/or slicing of data need to take place. A *Conditional Split* function divides the set of data into more subsets.
- **BI Transformation**
  Slowly changing dimension to track the changes and hold the historical information.

c. ***Load***
Once we have transformed the data into the required format, we load it into a destination data source. In our case, the destination is the data warehouse we designed to hold the pre-aggregated and transformed data.

Data Distribution. The output of the data analysis process is presented to the end users to assist them in making key business decisions. There are various tools available for presenting the information to business users. Microsoft Excel is one of the popular tools used in many organisation for disseminating data but its ability to support a rich presentation of data is limited and it does not provide a good support for data exploration, which is essential for this project as one of our main goals is to enabling self-servicing BI. In this project, we used the cloud-based version of Power BI [18] to present the results of our data analyses to company managers. We chose Power BI over the other self-service BI options available in the market mainly because Power BI is fairly easy to use even for non IT experts and it has the ability to integrate with Office 365 [22]. The skill set required to work with Power BI is very similar to that required to use Excel. With some basic training and documentation, our end users were easily able to generate and deploy reports and dashboards and share them with other users. In order to ensure data integrity, we made sure that the Power BI users can edit their own reports, but not the underlying data (Fig. 4).

Though Power BI was used to create and serve ad-hoc reporting requirements, there were some standard reports that should be delivered on a regular basis to the managers. Microsoft SQL Server Reporting Services (SSRS) [23] was used to create and deliver these standard reports. SSRS is a server based report generating software that comes as part of the Microsoft SQL Server suite. The SSRS provides a unique user interface based on Visual Studio that enables developers to connect to the relevant data but unlike Power BI, use of SSRS needed specialist technical knowledge and skills to create and distribute the reports and hence SSRS was not directly available to our end users. In this project, we used SSRS to create some standard KPI reports such as

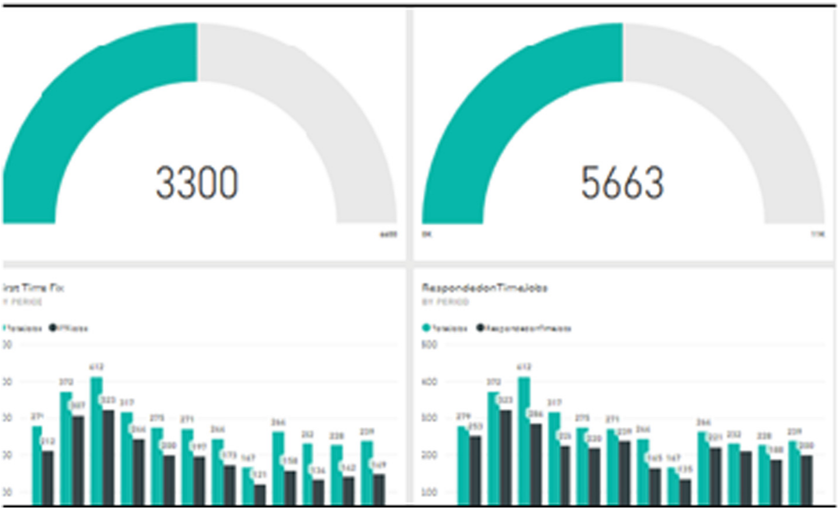


Fig. 4. Sample Power BI reports.

Engineer Performance and Productivity and Breakdown Calls on a monthly basis. These reports are embedded into an email which is sent to the respective managers for their analysis. As these are run-of-the-mill reports which can be automatically generated once the report generation has been set up, the extra IT staff time required is minimal. Our experience shows that enabling business users to explore the data and generate self-servicing business intelligence is a good approach as this reduces the communication overhead in generating BI reports. Once the business users have established the types of BI reports that are beneficial to their regular business decision-making process, the IT team can then implement these reports using SSRS and have them automatically delivered to the business users on a regular basis.

Feedback. In order for the BI solution to be effective, it is important to generate accurate and high-quality information. As developers we carried out unit testing to ensure our code is free from bugs. However, it was critical for the end users to verify the accuracy of the data and confirm the quality of the reports. Hence, the final step in each implementation iteration was to collect feedback from the end users. This feedback was based on the usability and accuracy of the data that was delivered to them. We made use of this phase to inform the next iteration of implementation. In this project, the feedback we obtained is from a steering group consisting of managers from various levels in the company who are also users of our BI solution. Our steering group members include commercial and technical directors as well as business operation managers.

3.4 Benefits

Based on this pilot implementation of a BI solution, we have noticed significant benefits for AGGORA in the first year of implementation. The key beneficiaries have been the top and middle level management. Using Power BI, the management teams have been able to analyse the performance by creating their own on-demand reports without depending on the IT team. This in turn has freed up the time spent by the IT team on creating reports and validating the information. We have estimated the time saving to be around 1.25 days per month.

The multiple benefits of implementing a BI solution in AGGORA can be classified into three categories:

i. *Improvements in Data Support*
This encompasses all the attributes related to reporting and its improvements. With the introduction of Power BI, it has been noted that more reports are being self-serviced than anticipated. The increase in the number of reports is mainly because of the reduction in overall effort involved in reporting and data analysis and the reports are being used in the business management. Power BI being more graphical than Excel has led to improved visualization of the data presented in the reports and also the flexibility to analyse results based on different dimension has accommodated even newer information needs.

ii. *Improvements in Decision Support*
This includes all the factors that are associated with better and more informed decision making. The on-demand reports using Power BI and standard reports using SSRS have been precise and present information that was previously not easily accessible, which, which has led to better informed business decisions. Furthermore, the ability to explore a larger subset of data without performance degradation or time lags has led to more timely business decisions being made and also facilitated the identification of business trends and the identification of risks. The use of a specialized data warehouse has also reduced the processing load of the live database, making the core business system more efficient, reliable and effective.

iii. *Savings*
The time and cost saving achieved by the introduction of the BI solution has been very evident. This implementation has led to a significant amount of time saving in AGGORA's IT department as there is no longer the need for the IT staff to produce reports using queries and pivot tables in an ad hoc manner. This has also opened up more opportunities for the IT staff to widen the scope of information available and improve their efficiency in other business areas.

Table 1 shows the tangible benefits achieved by AGGORA during the project development and the projected tangible benefits to AGGORA one year after the project completion.

Table 1. Tangible benefits.

	Tasks	Tangible benefits achieved during project	Tangible benefits projected for one year after project completion
1	Increased efficiency of AGGORA IT team	£3,220	£5,860
2	Introduction of Power BI for internal reporting	£4,202	£7,100
3	New reports built using data warehouse and Power BI to increase efficiency and revenue	£18,060	£60,000
4	Time saving achieved for IT team by using data warehouse for generating report	£7,000	£13,000
5	Improved performance of Asset management system	£1,000	£1,000
Total		**£33,482**	**£86,960**

4 Discussion

In this section, we will summarise our experiences in implementing the first iteration of a BI solution within a UK SME. As a first remark, we understood that it is vital for the BI solution to be user-friendly so that there would be more engagement from non-technical business users allowing them to better understand the benefits of using the solution. In this project, we have tried to abstract the technicality from end users by presenting the analytical results in the form of reports using SSRS and Power BI. The SSRS tool used to generate standard reports requires some technical skills to create and deploy reports. Despite of the technical complexity, the visually rich presentation of the data makes SSRS a powerful dissemination tool. We have noted that as Power BI is very similar to Excel, after some initial basic training, any competent Excel user should be able to use this tool with confidence and without much technical assistance. Power BI supports an intuitive process for creating on-demand reports and it also provides rich data visualisations. Furthermore, up-to-date tutorials demonstrating how to use Microsoft BI tools are readily available, making it highly appealing for end users.

It is important to understand the problem to be solved before implementing a BI solution. A good BI solution should focus on providing answers to important business-specific questions. In order for the business users to benefit from those answers, the Key Performance Indicators (KPIs) that will conform the metrics need to be established first. The implementation strategy must be clear and it will depend upon the business environment. In our project, we have used an iterative approach in order to ensure that the project meets the user's changing requirement over time.

As a part of implementation, it is vital to use a well-defined methodology to ensure an efficient BI solution. This requires technical knowledge of core BI concepts such as data warehousing, the ETL process and availability of different data visualisation tools. In terms of delivery strategy, we have chosen to deliver BI in a top-down manner, with the first set of BI solutions rolled out to top-level managers. This allowed top-level management to better understand the benefits of BI, and hence be more supportive of a wider exploitation of BI within the company.

Since BI implementation cannot be a one-off process, the design and implementation should always allow for changes to be made. Such a solution gets embedded within the organisation and starts to work with business. To keep the costs down and to make the solution more accessible to a wide range of end users, we have deliberately used only a limited set of BI tools. The Microsoft suite of BI tools was used for building all aspects of business intelligence required in this project. There are more tools offered in this Microsoft suite of BI application that can be used for more detailed data analytics such as prediction and forecasting.

Finally, our experience shows that:

An effective BI solution lets business users establish the performance metrics and measure their current performance against the KPIs.

While many SMEs understand the benefits of implementing BI solutions, they are often put off by the seemingly complex and expensive implementation process. Our experience shows that affordable and relatively simple BI solutions exist and they can easily be integrated into an SME's existing IT infrastructure.

A thorough understanding of the existing IT infrastructure is necessary to select and implement a custom BI solution within an SME.

Implementing a BI solution is a non-trivial process and involves several phases. A good understanding of these phases is needed to overcome some of the technical challenges. We have found that there is no need for training the entire team of IT staff up-front. Training up a single recent IT graduate is sufficient to gradually roll out the development process.

In order to make the BI solution successful, accurate intelligence needs to be delivered. It is important to have high-quality data which can achieved by identifying the data problems early and transforming and cleansing the data during the ETL process.

It is vital to use a well-defined methodology to design the data warehouse. In our project, we have used dimensional modelling approach to ensure efficient data retrieval and analysis.

The Microsoft suite of BI tools provides all the required components to implement an enterprise wide BI solution. This makes them perfectly suitable for SMEs wanting to implement BI solutions, who are already using other Microsoft business products as a part of their IT infrastructure.

While technical barriers do exist, there are government schemes, in the case of UK and Europe, available to help SMEs to overcome such barriers through funding, recruitment and knowledge transfer.

5 Future Work

This paper describes our first step in implementing a BI solution to an SME in the UK. Our next step includes using BI to: (1) achieve further potential enhancements within the organisation, (2) analyse the productivity of sub-teams within the enterprise and (3) perform margin analysis or implementing a more sophisticated account statement model.

With the option available to integrate R scripts with Power BI, we also plan to investigate using predictive time series analysis or other advanced data analysis to provide forecasts.

6 Conclusion

BI tools provide analytic data and key performance information which enables organisations of all sizes to be managed efficiently. It helps organisations to overcome the challenges involved in knowledge management and discovery. An efficient BI solution would potentially reduce the cost spent on resource and time to extract intelligence from the available data. It materializes the management's vision by empowering them with the ability to make more informed business decisions by minimizing the error on even large and complex data sample.

Based on our experience, we have understood that it is importance for SMEs to streamline their information resource in order to make more informed business decisions. We were able to appreciate the challenges that an SME could encounter while trying to implement a BI solution. In general, SMEs may not possess sufficient technical expertise that is needed to implement and maintain a custom BI solution. Furthermore, choosing the best solution from a densely populated analytics market is challenging and adds to the complexity.

In this case study, we have elaborated our approach to implement a BI solution for a UK based SME. Due to the volatile nature of the business within AGGORA, we opted for an iterative approach to implement the BI solution. It is important to establish the KPIs to lay the foundation for the implementation of the project. Considering the existing IT infrastructure, we decided to use Microsoft suite of BI tools to implement the BI solution.

We have also highlighted the need for using a sophisticated design approach to model the data warehouse. We have used the Kimball's approach for designing our data warehouse in order to ensure efficient and fast data retrieval. The source data was transformed and cleansed using the Microsoft integration services. The transformed data is presented to business users using a range of data presentation tools.

One of the main objectives of this project was to empower the business users with all relevant information to make more informed business decisions. We have delivered business intelligence using self-service BI and also standard reports. Self-service BI delivered using Power BI empowers managers to explore the available data and generate their own reports and dashboards without any involvement from the IT team. This has made the IT team more efficient, leading to cost saving from no longer having to generate ad-hoc reports. The standard reports delivered on a regular basis using

Microsoft reporting services helped the management team to identify patterns in performance transaction data which leads to improvements in the efficiency and productivity of the company and better informed management decision making.

We appreciate the possible limitations of our work. As this case study is based on a company which has already adopted a Microsoft-based IT infrastructure, we are mindful that the costs, benefits, implementation approach and time scales could be different for companies using different IT products within their infrastructure. However, with Microsoft being the market leader in both operational database management systems and BI and analytics platforms [7, 8], our findings are relevant to a large number of SMEs.

Acknowledgements. The project reported in this case study is funded by Innovate UK via the Knowledge Transfer Partnership Scheme (KTP009714). We acknowledge the financial support by Innovate UK. We also acknowledge the advice provided by our KTP Adviser Ms Jose Freeman throughout the project development. We thank the team at AGGORA who made this project possible through their participation.

References

1. AGGORA Group: AGGORA. http://aggora.co.uk/ (2016). Accessed 8 Mar 2017
2. Burstein, F., Holsapple, C.: Handbook on Decision Support Systems. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-48713-5
3. Chugh, R., Grandhi, S.: Why business intelligence? Significance of business intelligence tools and integrating BI governance with corporate governance. Int. J. e-Entrep. Innov. **4**(2), 1–14 (2013). http://www.igi-global.com/article/why-business-intelligence/89282
4. European Commission: What is an SME? http://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition/index_en.htm (2016). Accessed 8 Mar 2017
5. Fink, K., Ploder, C.: Knowledge management toolkit for SMEs. Int. J. Knowl. Manag. **5**(1), 46–60 (2009)
6. Gartner Inc.: Gartner Executive Programs' Worldwide Survey of more than 2,300 CIOs shows flat IT budgets in 2012, but IT organizations must deliver on multiple priorities (2012). http://www.gartner.com/newsroom/id/1897514. Accessed 07 Mar 2017
7. Gartner Inc.: Magic quadrant for operational database management systems (2015). https://www.gartner.com/doc/reprints?id=1-2PMFPEN&ct=151013. Accessed 15 Aug 2016
8. Gartner Inc.: Magic quadrant for business intelligence and analytics platforms (2016). https://www.gartner.com/doc/reprints?id=1-2XXET8P&ct=160204. Accessed 15 Aug 2016
9. Google: Google Charts (2016). https://developers.google.com/chart/. Accessed 8 Mar 2017
10. Horakova, M., Skalska, H.: Business intelligence and implementation in a small enterprise. J. Syst. Integr. **4**(2), 50–61 (2013)
11. Howson, C.: Successful Business Intelligence. McGraw-Hill, New York (2008)
12. InnovateUK: Knowledge transfer partnerships (2016). https://connect.innovateuk.org/web/ktp. Accessed 7 Mar 2017
13. Kimball, R., Ross, M.: The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd edn. Wiley, Indianapolis (2013)
14. Lacko, L.: Business Intelligence v SQL Serveru 2008. Computer Press, Brno (2009)
15. Luhn, H.P.: A business intelligence system. IBM J. Res. Dev. **2**(4), 314–319 (1958)

16. McCabe, L.: Closing the business intelligence gap for small businesses (2012). https://lauriemccabe.com/2012/01/27/closing-the-business-intelligence-gap-for-small-businesses. Accessed 8 Mar 2017
17. McGonagle, J.: An examination of the 'classic' CI model. J. Compet. Intell. Manag. **4**(2), 71–86 (2007)
18. Microsoft: Frequently asked questions about Power BI Microsoft Power BI (2015). https://powerbi.microsoft.com/en-us/documentation/powerbi-frequently-asked-questions/?CorrelationId=a0b3a797-a620-4e09-b346-1141a5f1d03d&ui=en-US&rs=en-US&ad=US. Accessed 8 Mar 2017
19. Microsoft: DAX queries (2016). https://msdn.microsoft.com/en-us/library/gg492201.aspx. Accessed 7 Mar 2017
20. Microsoft: Editions and components of SQL Server 2012 (2016). https://technet.microsoft.com/en-us/library/ms144275%28v=sql.110%29.aspx. Accessed 8 Mar 2017
21. Microsoft: Microsoft Developer Network (MSDN) (2016). https://msdn.microsoft.com/. Accessed 8 Mar 2017
22. Microsoft: Office 365 (2016). https://products.office.com/en-gb/business/explore-office-365-for-business. Accessed 8 Mar 2017
23. Microsoft: Reporting Services (SSRS) (2016). https://msdn.microsoft.com/en-us/library/ms159106.aspx. Accessed 8 Mar 2017
24. Negash, S., Gray, P.: Business intelligence. In: Proceedings of the Ninth Americas Conference on Information Systems (AMCIS 2003), Tama, FL, pp. 3190–3199 (2003)
25. Papazoglou, M.P.: Service-oriented computing: concepts, characteristics and directions. In: Proceedings of the Fourth International Conference on Web Information Systems Engineering, (WISE 2003), Rome, pp. 3–12 (2003). https://doi.org/10.1109/wise.2003.1254461
26. Raj, R., Wong, S.H.S., Beaumont, A.J.: Business intelligence solution for an SME: a case study. In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), KMIS, vol. 3, pp. 41–50. SCITEPRESS – Science and Technology Publications, Lda (2016). ISBN 978-989-758-203-5
27. Scholz, P., Schieder, C., Kurze, C., Gluchowski, P., Böhringer, M.: Benefits and challenges of business intelligence adoption in small and medium-sized enterprises. In: 18th European Conference on Information Systems (2010)
28. Tutunea, M.F., Rus, R.V.: Business intelligence solutions for SME's. Proc. Econ. Finance **3**, 865–870 (2012)
29. Watson, H., Wixom, B.: The current state of business intelligence. Computer **40**(9), 96–99 (2007)

Congestion Control Supported Dual-Mode Video Transfer

Juha Vihervaara(✉), Teemu Alapaholuoma, Tarmo Lipping, and Pekka Loula

Pori Campus, Tampere University of Technology, Pori, Finland
juha.vihervaara@tut.fi

Abstract. Transfer of videos over the Internet has increased considerably during the past decade and recent studies indicate that video services represent over half of the Internet traffic, with a growing trend. For the user-friendly operation of the Internet, it is important to distribute these videos in a proper and efficient way. However, no congestion control mechanism suitable and widely used for all kinds of video services is available. We have developed a congestion control mechanism, which is particularly suitable for long-living video traffic. The advantage of the proposed mechanism is its dual-priority nature. There is a mode for low priority traffic where the bandwidth is given away to other connections after the load level of a network exceeds a certain level. On the other hand, the real-time mode of the mechanism acquires its fair share of the network capacity. The real network tests of this study verify the proper operation of our congestion control mechanism.

Keywords: Congestion control · Video transfer

1 Introduction

Transfer of videos over the Internet has increased considerably during the past decade. Cisco forecasts that Internet video traffic play a big role also in the future [1]. It predicts that video traffic will form 82% of all consumer Internet traffic by 2020. Internet video to TV will continue to grow at a rapid pace, increasing 3.6-fold by 2020. Virtual reality based applications will also increase the video type traffic of the Internet. Videos are widely used because video-based solutions offer advantages and possibilities for many application areas. For example, in education, the use of video-based instructional materials often produce better learning results compared to the traditional print-based materials [2]. In addition, YouTube can be considered as an important tool for education [3].

Due to high popularity of video traffic, it is also an important cause of network congestions. Network operators have largely relied on overprovisioning and TCP congestion control to avoid congestions in their networks. However, unnecessarily high overprovisioning with high power consumption does not promote green Internet ideology [4]. Although some video services use TCP to implement their transport services in a manner that actually works, TCP's transport service is not suitable for all video applications. By implementing retransmissions, TCP offers reliable transport services

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 326–345, 2019.
https://doi.org/10.1007/978-3-319-99701-8_16

to applications. Normally, a real-time video application does not require retransmissions because this type of applications are often loss-tolerant. Occasional packet drops do not degrade the quality of service experienced by the users of these applications. These packet drops can be alleviated by using the error correction properties of the applications. If the application is working in a real-time repeat mode, the order delivery property of TCP may cause problems. Due to the head-of-the-line blocking problem, the bytes following the missing ones cannot be delivered to the application. TCP's bursty-like transmission also causes delay jitters and sudden quality degradations because there can be abrupt and deep sending rate reductions. For these reasons, real-time video applications often prefer to use the unreliable UDP protocol. Unfortunately, UDP does not implement congestion control.

The approach of using congestion control only with TCP traffic has been appropriate in the past because TCP has represented major proportion of network traffic. However, nowadays UDP based long living communication events are common due to the popularity of various video services among consumers. It makes sense to equip these communication events with congestion control. This may offer new opportunities for old and new congestion control mechanisms to become deployed.

There are different kinds of ways to use video over the Internet. With live broadcasting, only a moderate buffering can be used at the receiver side due to the real time requirements. Therefore, delay requirements and bandwidth demands are important. On the other hand, in non-real-time applications where extensive buffering can be utilized at the receiver side and, therefore, delay and bandwidth demands are not important, some kind of background loading may be preferred. For example, the service provider can download content to proxy servers by using backward loading. The case can also be some kind of intermediate form. At first, the video can be transferred with the high speed. When there is enough data in the receive buffer, the transfer mode can be switched to the backward loading type.

So, two different kinds of transfer modes are needed in modern video services: a backward loading mode where delay and bandwidth demands are moderate and a real-time mode where delay and bandwidth demands are of high priority. Based on these different kinds of demands, the two modes also need different kinds of congestion control mechanisms. The backward loading mode has to work like a low-priority service in which the bandwidth is given away to other connections when the load level of the network is high enough. In contrast, the real-time mode always wants its fair share of the bandwidth.

Many congestion control mechanisms have been developed to be used either with low priority or with real-time services. However, little research effort has been put into developing a mechanism suitable for both modes. We recently developed this kind of integrated mechanism that supports both of these transfer modes. This mechanism was named Congestion control for VIdeo to Home Internet Service (CVIHIS). The algorithm was presented in the paper [5], where CVIHIS's performance was analyzed by extensive simulations. In the paper [6], we tested the operation of CVIHIS in real network environments. In these real network tests, CVIHIS was tested against itself more comprehensively than in the simulations. This paper re-presents and refines the results of the paper [6]. This study takes into account the situations where two congested routers simultaneously occur on the transmission path. The paper also analyzes how CVIHIS will work with the common problems of delay-based congestion control mechanisms.

The paper is organized as follows: Sect. 2 outlines congestion control backgrounds; Sect. 3 introduces our dual-mode congestion control algorithm fir video services; Sect. 4 presents the test results and Sect. 5 concludes the paper.

2 Congestion Control

Congestion control principles of the Internet are presented in this section. Congestion control is a wide research area and only issues relevant for this study are introduced here.

2.1 Importance of Congestion Control

In its basic form, the Internet is built upon an assumption of best effort service. This means that the network does its best to deliver data packets to receivers as quickly as possible. On the other hand, the best effort principle also means that the network does not guarantee anything. It is not against Internet's laws that packets are queued or dropped inside the network. Congestion situations handled by queuing and dropping of packets are therefore fully acceptable. Of course, from a network user's point of view, congestion is not a desired situation. Therefore, a network claiming to operate in a user friendly manner must implement some kind of congestion control.

If congestion control is implemented in an inoperative way, serious troubles may occur. When some part of the network is in a congested state, it queues traffic and packets may be dropped. Therefore, receivers do not receive the expected packets in time and senders cannot get acknowledgements inside the time limits. After that, senders, which are implementing reliable communication, will start to resend packets causing further congestion. This can lead to congestion collapse in which case only little useful communication is happening through the network [7].

Many reasons can lead to a congested network. The paper [8] specifies such kind of reasons: limited capacity of the routers, load of the network, link failures, heterogeneous bandwidths. The consequences of inoperative congestion control are discussed in [9]. They point out that in a congested network large queuing delays are experienced, which increases the response times of web services.

The background of congestion control is in queuing theory [10] as packets move into and out of queues when they pass through a network. Therefore, packet-switched networks can be considered as networks of queues. However, it is good to remember that extensive queuing inside the network is not a desired operation. Queue lengths should not reflect the steady condition we want to maintain in the network. Instead, they should reflect the size of bursts we need to absorb [11]. The goal of congestion control is to avoid a congestion situation in network elements. By another, more sophisticated definition, the target of congestion control is to adapt the sending rates of senders to match the available end-to-end network capacity. This definition emphasizes the fact that network-wide approaches must be used to implement congestion control. Otherwise, congestion is only shifted from one node to another. Therefore, in theory, we should monitor traffic in the whole network.

2.2 Congestion Control Mechanisms for Video Services

Because the TCP and UDP protocols are not completely suitable for video services, there is the need for a protocol that takes into account the requirements of video traffic. In this section, some congestion control algorithms, suitable for video traffic, are presented. Each of these algorithms is suitable for either a low priority or a real-time service. None of them has been developed with both these service types in mind.

LEDBAT [12] is designed for low priority applications. It has been used in some background bulk-transfer applications such as BitTorrent, for example. It provides low priority services by using one-way delay measurements to estimate the amount of queued data on the data path. When the estimated queuing delay is less than the predetermined target, LEDBAT concludes that the network is not yet congested and it increases its sending rate to utilize the free capacity of the network. When the estimated queuing delay becomes larger than the predetermined target, LEDBAT decreases its sending rate as a response to the potential congestion. The sending rate is increased and decreased more aggressively if the queuing delay is far from the target. TCP-LP [13] is another delay-based congestion control protocol for low priority services.

The next two algorithms are suitable for the real-time mode as they want their fair share of the bandwidth. The best known proposal for video services is DCCP [14] and its TCP Friendly Rate Control version [15], abbreviated as TFRC. DCCP offers congestion control for UDP-like unreliable applications. DCCP can be briefly described as TCP without byte-stream semantics and reliability, or as UDP with added congestion control, handshakes and acknowledgments for congestion feedbacks. The main issue with DCCP's congestion control is that the congestion control is not a part of DCCP itself but DCCP allows applications to choose from a set of congestion control mechanisms. Therefore, different kinds of congestion control mechanisms can be used with DCCP, TFRC being one of them. TFRC uses a throughput equation to calculate the allowed sending rate. Because DCCP tries to be fair against TCP, it is natural that TFRC uses the TCP throughput equation. TFRC is designed for applications that require smooth rate. Therefore, TFRC responds to the changes of the available bandwidth more slowly than TCP.

Google Congestion Control for Real-Time Communication [16] is a new proposal in this area. It defines two congestion control methods: one for the sender side and another for the receiver side. Either both or only one of these methods can be used. The receiver side uses delay gradients in a sophisticated way to detect congestions. The sender side method is based on information about round-trip times and packet losses.

One possibility to achieve a dual-mode congestion control mechanism, such as the one presented in this study, would be to put together the best low-priority and real-time congestion control algorithms. However, this kind of implementation would be ungainly, especially when the mode has to be changed on the fly. The real dual-mode mechanism presented in this study allows the change between the modes in a seamless way.

2.3 TCP Friendliness

The real-time mode of CVIHIS aims to share the bandwidth of transmission links in an equitable manner. This equal allocation of bandwidth is called friendliness. Often the term TCP friendliness is used as in the past years most of the traffic flows were TCP flows and the TCP protocol has traditionally been responsible for the congestion control of the Internet. Therefore, it is a natural choice to compare a new mechanism against the TCP protocol. The basic idea is to protect existing TCP flows from the flows that use too aggressive congestion control mechanisms.

Unfortunately, TCP friendliness is a complicated concept. Even a TCP flow itself is not always friendly against another TCP flow. Several versions of the TCP protocol exist and these versions are not completely identical in their behaviours. In addition, TCP's throughput degrades in case of higher round-trip times (RTT) [17]. Therefore, TCP has a bias against high-RTT connections giving preference to the users with short RTTs. Several improvements such as the Delayed ACK mechanism [18], for example, have been suggested to make TCP congestion control work in a better way. Unfortunately, only some TCP implementations have adopted these improvements and, therefore, different code implementations behave in different ways. Due to this, even identical TCP implementations are not equal. Another problem is that there is no exact definition for the concept of TCP friendliness. When a new mechanism is developed and compared against the TCP protocol, there is always some room for personal opinions.

3 Dual-Mode Congestion Control Mechanism CVIHIS

The algorithm of CVIHIS is introduced in this section with brief description of the implementation principles of CVIHIS for real network tests.

3.1 Basic Properties of the Algorithm

CVIHIS is a receiver-based mechanism so that most of the processing can be done at the receiver side instead of the heavy loaded server side. It complies with the end-to-end approach, which states that complex issues should not reside in routers. Because the sending rates of video applications should usually vary in a smooth way, CVIHIS is a rate-based congestion control approach. The window-based control is seldom suitable for continuous multimedia streaming because it tends to produce bursty-like traffic behavior [19]. Exponentially Weighted Moving Average (EWMA) is also used by CVIHIS to filter out quick rate changes.

If the network does not deliver explicit congestion feedbacks, the sending rate adjustment can only be based on packet losses or delays. Both indicators are utilized by CVIHIS, but the algorithm is somewhat more delay-based than loss-based. The reason for emphasizing the delay-based approach is that it generates suitable conditions for implementing the low-priority behavior [20]. CVIHIS uses one-way delays in delay measurements so that the necessary conclusions can be made at the receiver side. However, using one-way delays also has other benefits that are explained by the paper [21].

3.2 Backward Loading Mode

Figure 1 presents the rate adaptation schema of the backward loading mode. The algorithm of CVIHIS keeps track of two delay values, minDelay and maxDelay, based on one-way delay measurements. The minDelay value corresponds to the situation when the queues of the routers are empty on the connection path. The minDelay value includes only propagation delay components, not queuing delays. The minDelay value is the shortest delay value experienced during the lifetime of the connection. Instead, the maxDelay value includes the queuing delay component. It corresponds to the situation in which the buffer of a router overflows. Therefore, maxDelay is updated every time when a packet drop occurs. CVIHIS uses the delay value of the last received packet prior to the dropped packet for the maxDelay value.

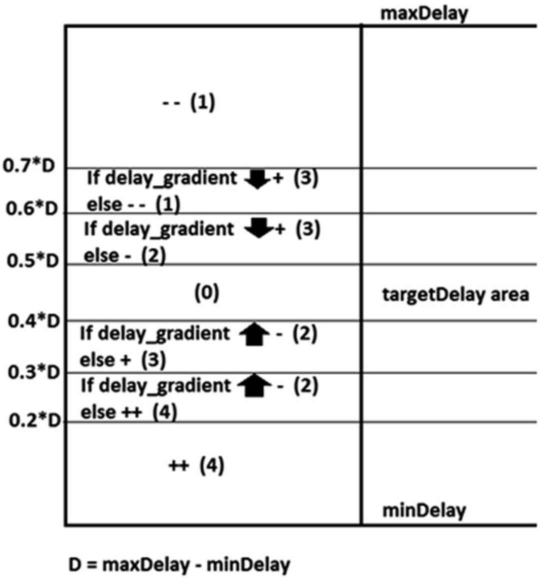


Fig. 1. Rate adaptation schema of CVIHIS (Source: [5])

With the help of these two delay values, the delay space is divided into seven rate adaptation areas. It could also be said that the queue of the router is divided into several corresponding parts. The seven rate adaptation areas are used so that sufficiently accurate information about the state of the network can be provided to end-hosts. CVIHIS' objective is that it tries to keep the queue at the level of the target delay area. When operating in the upper delay areas, CVIHIS decreases its sending rate and, when operating in the lower delay areas, CVIHIS increases its sending rate. The positioning factors for each delay area are presented on the left side of Fig. 1. The targetDelay area is not placed in the middle of the delay space but is shifted somewhat downwards so that queues can be kept short.

The black arrow inside some of the delay areas in Fig. 1 represents the delay gradient obtained by comparing the delay values of two consecutive packets. If the arrow points upwards, delays are increasing, delay gradient is positive, and the queue is filling. If the arrow points downwards, delays are decreasing, delay gradient is negative, and the queue is emptying. Inside the four delay areas with the arrows, the rate adaptation command is based on the actual delay value and the value of the delay gradient. The rate adaptation scheme tries to achieve two targets: it tries to drive the queue level to the target delay area by measuring the actual delay value and, on the other hand, it tries to adapt the sending rate according to the bottleneck capacity. This is done by means of the delay gradient. If there is a conflict between the delay area and delay gradient adaptation, the gradient adaptation is chosen. The two extreme delay areas do not use delay gradients for rate adaptation decisions because these areas are far away from the targetDelay area.

In its additive increase phase, the TCP protocol increases its sending rate by one segment for each round-trip time interval. In its basic form, CVIHIS increases or decreases its sending rate by one packet for each square root of a round-trip time interval. By using square root, CVIHIS alleviates the favoring behavior of short distance connections.

CVIHIS adjusts its sending rate through seven adjustment steps. Six of these steps are presented in Fig. 1. Bigger steps are used when the queue level is further away from the target. In Fig. 1, the step sizes are denoted by different number of + or − marks. If there are three marks, CVIHIS increases or decreases its transmission rate by one packet for each square root of the round-trip time. If there are two marks, the adjustment steps are smaller. The smallest steps are indicated by one + or − mark. To enter the targetDelay area in a smooth way, CVIHIS uses short steps in the delay areas just beside the targetDelay area (rate adaptation feedbacks 2 and 3). The adjustment steps related to the delay gradients (rate adaptation feedbacks 6 and 7) are the shortest ones. The seventh adjustment step is a multiplicative decrease step taken after a packet drop. The multiplicative decrease step is taken only once per a round-trip time cycle.

Table 1 presents the rate adjustment factors of CVIHIS. These factors are set so that CVIHIS can compete in a fair manner with the TCP NewReno version. The integer values in brackets refer to the rate adjustment commands of CVIHIS presented in Fig. 1. All decision procedures related to Fig. 1 are implemented at the receiver side. Only the rate adaptation commands are transmitted to the sender. In the case of four leftmost columns, the rate adjustment is based on the square root of the round-trip time. The factor expresses how many more or less packets will be sent during the next square root of the round-trip time than just before. MD is a multiplicative decrease factor used after packet drops to increase the sending gap of packets. SF is a smoothing factor used for the EWMA filter to filter out quick rate changes. PF is a pushing factor used only by the real-time mode.

Table 1. Adjustment parameters of CVIHIS (Source: [6])

−−− (1) +++ (4)	−− (2)	++ (3)	− (6) + (7)	MD (5)	SF	PF
1.0	0.7	0.5	0.2	1.10	0.5 * last update 0.5 * history	1.05

3.3 Real-Time Mode

The backward loading mode backs off when it competes with TCP. In order to be suitable for the real time mode, the implementation code has to be modified so that it will behave in a more aggressive way. On the other hand, it is desirable that the code implementation of the backward loading mode is modified as little as possible. Both of these goals can be achieved in a simple way by using an approach in which the minimum delay value is pushed upwards in a continuous manner. This means that the delay areas of Fig. 1 are also pushed upwards and, therefore, CVIHIS behaves in a more aggressive way. Shifting the delay areas upwards is only done when competing behavior is actually needed. If the last measured delay value is smaller than the pushed minimum delay value, the minimum delay value is set to the value of the last measured delay.

This kind of minimum delay pushing means that the real-time mode of CVIHIS is not a delay-based congestion control solution any more. The pushing operation shifts this version towards loss-based congestion control. Therefore, the real-time mode of CVIHIS is a kind of hybrid solution, a delay-loss-based solution. The minimum delay value is pushed upwards in a multiplicative way. It was found that the pushing factor of 1.05 is suitable.

It is worth noting that CVIHIS is a pure congestion control mechanism. If the application is delay sensitive, delay requirements must be satisfied by Quality-of-Service mechanisms [22]. The dual mode mechanism presented in this study can also be achieved using Quality-of-Service techniques. In this case, the mechanism could be called a dual priority mechanism. However, this kind of mechanism can not be a pure end-to-end mechanism as the implementation would require network support at least to some extent.

3.4 Software Implementation

The code implementation of CVIHIS should be placed somewhere in the protocol stack to enable real network tests. There are several possibilities for this placing. For example, the kernel implementation of an open source operating system could be modified. In this way, the UDP implementation of the operating system could be adjusted to correspond to CVIHIS' algorithm. Instead of using this kind of elaborate solution, an easier implementation option was chosen. CVIHIS was implemented through a normal socket program on top of the UDP protocol. This solution is possible because the UDP protocol does not offer any special transport services, which could disturb the operation of CVIHIS. Unlike in simulation environment, the real network implementation should take into account that the clocks of the source and receiving ends have not been synchronized with each other. Therefore, the real network version has to obtain round-trip times by actual measurements. When a packet is transmitted, the sender side stores the transmission timestamp. When the corresponding acknowledgment arrives, the round-trip time is calculated.

In addition, this implementation option provides further advantages. Same computer can be used to run a few traffic sources in parallel because these traffic sources can be separated from each other by means of UDP port numbers. This reduces the number of computers needed for the test network. The solution also offers resistance against firewall blocking if large scale real network testing is done in the future.

4 Test Results

In this section, the structure of the test network is described and the most relevant results of the network tests are presented. This study presents the results related to the paper [6] in an extended way. The test results of CVIHIS shows that there is room for improvements in some cases. In these cases, there can be present connections which posses very different round-trip times. As it is mentioned, TCP favors the connections of short round-trip times, therefore, it is necessary to adjust the algorithm of CVIHIS to favor short round-trip times as well. However, CVIHIS performs this favoring in a more moderate way than TCP. In fact, it is not necessarily a bad idea to favor the connections of shot round-trip times. Often these short connections consume less network resources than long way connections. By favoring connections of short round-trip times, network operators can maximize the total traffic volume in their networks. The challenge is to find a suitable balance so that fairness does not suffer too much.

4.1 Test Network

Figure 2 presents the structure of the test network. There are four end nodes and three routers. The links between the routers form the bottleneck links of the connection path. With this simple test network structure, and with the help of the tc (traffic control) program, it is possible to emulate different types of networks. Tc [23] is the Linux utility program used to configure the kernel packet scheduler.

Tc is utilized in two ways to vary the configuration of the test network. Tc is used to control the traffic in the Linux routers. In this way, the capacity of the bottleneck link and the queue size of the link can be varied. When traffic is controlled, the transmission rate of the link is under control. Typically, this means that the available bandwidth is decreased. Traffic control can also be used to smooth the burstness of incoming links by defining the queue size of the link. If the queue size is exceeded, the incoming packets are dropped. At the end nodes, tc is used to define the delay characteristics of the outgoing links. In this way, it is possible to emulate different round-trip times of connection paths.

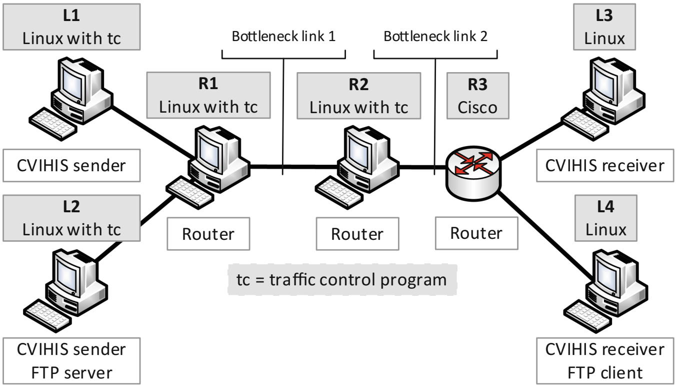


Fig. 2. Structure of the test network.

4.2 Backward Loading Mode

The goal of the backward loading mode is to achieve stable sending rate if there is no need for the backoff function. This was ensured by performing twenty tests in the test network. In the test cases, the queue size of the bottleneck link varied between 40 and 60 packets, the capacity of the bottleneck link varied between 2 and 4.5 Mbps and the round-trip time varied between 10 and 250 ms. The sending rate stabilized in all cases.

Another objective of this mode is proper backoff behavior. This was verified against one TCP NewReno connection by using the same kinds of test setups as in the case of the stability check. The proper backoff behavior was observed in all cases. In Fig. 3, the sending rates of the CVIHIS connections are presented in test cases, where the backoff behavior was verified by using three different round-trip times (10, 100 and

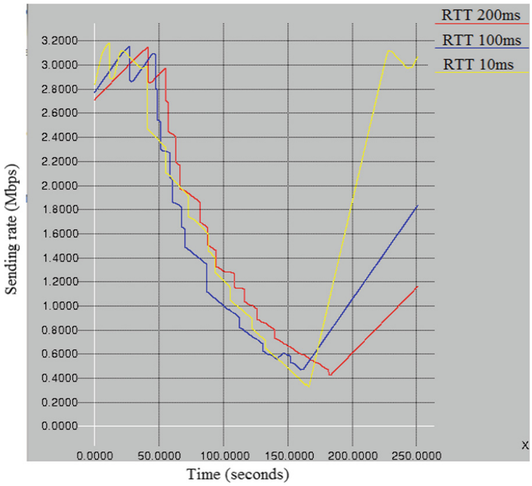


Fig. 3. Simulation results of the backward loading (Source: [6])

200 ms) for CVIHIS. TCP used the round-trip time of 200 ms in all these cases. The capacity of the bottleneck link was 3 Mbps. The TCP connections were active between the test time of about 50–170 s. As it can be seen, CVIHIS increases its sending rate faster if the round-trip times are short.

4.3 Real-Time Mode

The TCP-friendliness of CVIHIS was tested against the TCP NewReno version by performing twenty six tests. The queue size of the bottleneck link was 50 packets and the capacity of the bottleneck link varied between 2 and 4.5 Mbps. Four different round-trip times (20, 80, 140 and 200 ms) were used. The starting rates of the connections also varied among the tests.

Figure 4 presents the test results. As it can be seen, individual measurements depart from the trend due to the phase effect. The figure presents the proportion of the CVIHIS connection from the capacity of the bottleneck link. The figure indicates acceptable level of averaged fairness. In the worst case, the connection of higher bandwidth gets about 1.6 times as much bandwidth as the slower connection.

As mentioned earlier, TCP favors the connections of short round-trip times while CVIHIS does this in a more modest way. The above results confirm this. Round-trip times affect CVIHIS less than TCP. CVIHIS manages relatively modestly when round-trip times are short. When round-trip times are long, CVIHIS manages somewhat better than TCP.

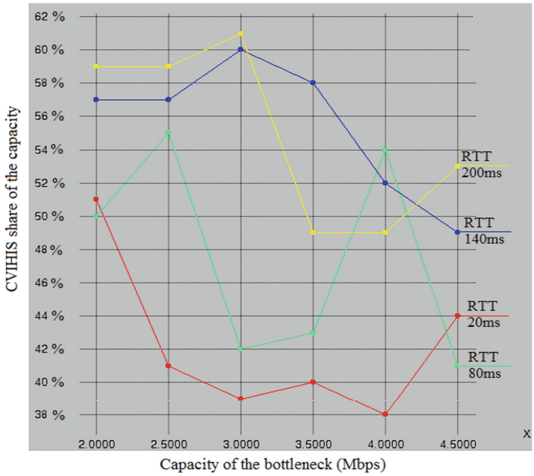


Fig. 4. Real-time mode against one TCP connection (Source: [6])

The Linux version used in the real network tests also supports another TCP congestion control mechanism. This version is CUBIC TCP [24]. In fact, CUBIC is the current default TCP algorithm of Linux. Therefore, CVIHIS was also tested against the

CUBIC version. The preliminary results show that CUBIC behaves somewhat more aggressively than NewReno. If CVIHIS is desired to manage in a friendly way against the CUBIC version, the rate adjustment parameters of CVIHIS have to be adjusted slightly so that CVIHIS would behave more aggressively.

4.4 CVIHIS Against Itself

The results of the previous subsection and the paper [5] show that it is challenging to attain acceptable level of fairness in heterogeneous network environments. Hence, implementing a well-performing solution for network congestion control might require that there are only a few kinds of congestion control mechanisms on the Internet. Thus, it is important that CVIHIS behaves in a fair manner also against itself.

The real-time mode of CVIHS was tested against itself by doing 30 tests. In the test cases, the queue size of the bottleneck link varied between 40 and 60 packets, the capacity of the bottleneck link varied between 2 and 6 Mbps, and the round-trip time of the connection path varied between 10 and 240 ms. In some of the tests the connections used different round-trip times. Also, the starting rates of the connections varied among the tests. Based on these tests, the sending rates indicate good level of fairness. In most cases, transmission rates differed less than 10%. Only when round-trip times were significantly different, the rate differences were larger than 10%. In the worst case, the connection of higher bandwidth got about 1.7 times as much bandwidth as the slower connection. In this case, the faster and slower connections had the round-trip times of 10 and 180 ms, respectively.

Figure 5 presents the result of one of the tests. The capacity of the bottleneck link is 4 Mbps and the round-trip times of the connections are 60 (red) and 180 (blue) milliseconds. In this case, the average sending rates are 2.085 Mbps and 1.935 Mbps. The first 50 s where omitted when calculating averaged rates.

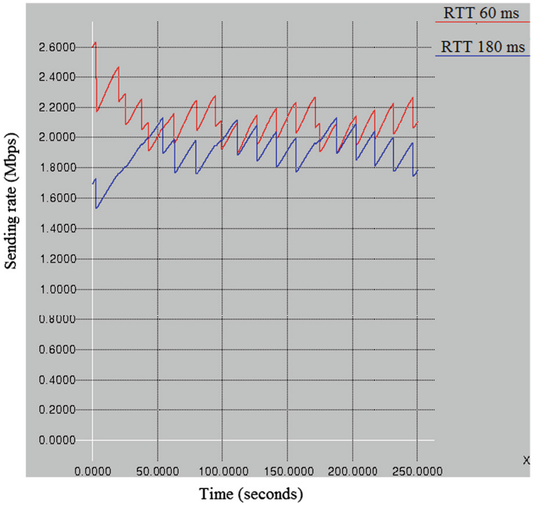


Fig. 5. CVIHIS real-time mode against itself (Source: [6])

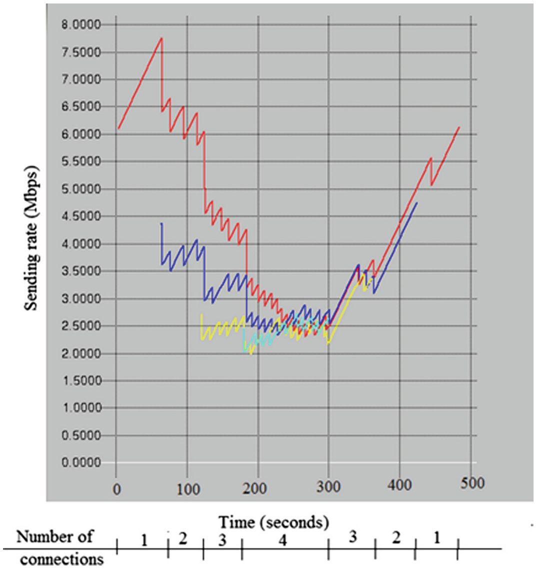


Fig. 6. Four real-time mode connections (Source: [6])

Some tests were made involving four CVIHIS connections in the active state at the same time. CVIHIS performed in an acceptable way in these tests although it took more time to balance the sending rates when round-trip times were long. Figure 6 presents the result of one such test. In this case, the capacity of the bottleneck link was 10 Mbps and the queue size of the bottleneck link was 60 packets. The round-trip time was 30 ms. The number of connections in the active state is shown at the bottom part of this figure.

The backoff behavior of CVIHIS in the backward loading mode was also tested when there was a real-time connection on the connection path at the same time. Fifty tests were made so that the capacity of the bottleneck link was 2 or 4 Mbps and the queue size of the bottleneck link was 60 packets. The round-trip times varied from 10 to 200 ms.

When both modes had the same round-trip time, the backoff action was as expected. Tests were also performed using different round-trip times for the modes. In these tests, the backoff action occurred slowly if the real-time mode connection had significantly longer round-trip time than the backward loading mode connection. When the round-trip times differed considerably, ten times, for example, backoff action did not take place at all. Figure 7 illustrates the above-mentioned behavior. In this figure, the sending rate of the backward loading mode is presented in three separate cases. The round-trip times of the real-time mode connection are 10, 150, and 200 ms in these cases. In all these cases, the round-trip time of the backward loading mode connection is 50 ms. It is fairly easy to moderate this phenomenon by adjusting the parameters of the backward loading mode so that it would behave less aggressively than the real-time mode. In this study, both modes shared the same parameter set.

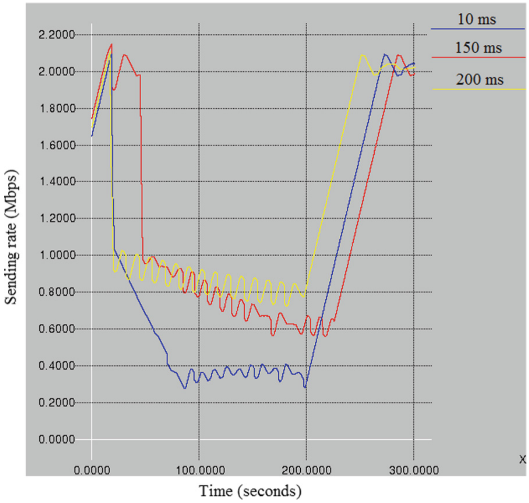


Fig. 7. Backoff behavior of the backward loading mode (Source: [6])

4.5 Case of Two Queues

There has been only one non-empty queue on the connection path in the previous test cases. As far as this single non-empty queue condition is met, the behaviour of our test network structure is compatible with that of more complicated network structures. The location of this non-empty queue can change if the size of the queue and the bandwidth of the out-going link remain similar. The more complicated network structures have been taken into account by using different one-way propagation delay values in the previous test cases. However, in real networks, it can happen that there are several non-empty queues on the connection path at a certain moment. In this subsection, the case in which there are two non-empty queues on the connection path is tested. The case of two non-empty queues affects especially the maximum delay value so that it is not static any more.

Now there are three end nodes that send traffic by using the real-time mode of CVIHIS. The third source is located in Linux Router 2. The receiver of this third connection is located in the receiving hosts L4. This host runs two receiving processes at the same time. There are now two bottleneck links which reside in the Linux routers. The connection R2-L4 has only one bottleneck while the other two connections have two bottlenecks. We want to test the case in which there are occasionally two non-empty queues on the connection path. So, the capacity of the second bottleneck link should be 1.5 times as much as the capacity of the first bottleneck link, or a little bit more. This is because the first link has two connections and the second link three connections.

Six tests were made to study CVIHIS' fairness against itself. The queue sizes of the bottleneck links were 40 packets and one-way propagation delays were 50 ms in all the cases. The results of these tests are presented in Table 2. The second and third columns present the capacity of the bottleneck links. The actual test results are presented in the last three columns. These columns present the sending rates of the connections and the standard deviations of CVIHIS' sending rates. The standard deviations are presented inside the parentheses. Based on these results, it can be said that the sending rates indicate good level of fairness. We also carried out tests, in which one of these three connections owned the one-way propagation delay value of 150 ms. These tests also indicated good level of fairness. In the worst cases, the faster comparable connection gets about 1.15 times as much bandwidth as the slower connection.

Table 2. Ten tests for testing CVIHIS' friendliness with two queues.

	Capacity of link R1-R2 (kbps)	Capacity of link R2-R3 (kbps)	CVIHIS 1 L1-L3 (kbps)	CVIHIS 2 L2-L4 (kbps)	CVIHIS 3 R2-L4 (kbps)
1	4000	6000	2008 (33)	2050 (32)	–
2	4000	6000	2005 (47)	2045 (42)	2040 (53)
3	4000	6500	1929 (76)	2005 (86)	2646 (75)
4	6000	9000	3049 (58)	3053 (57)	–
5	6000	9000	3047 (82)	3042 (87)	3048 (60)
6	6000	10000	3081 (83)	3014 (84)	3968 (168)

The results indicate that CVIHIS' sending rate varies more in the case of two queues than in the case of one queue. This can be seen when comparing the results of the rows 1 and 2 to each other. The same is also true for the rows 4 and 5. This is because the maximum delay value related to the packet drop situations is not static any more. The maximum delay value varies according to the level of the non-full queue. This CVIHIS' sending rate fluctuation can also be seen in Fig. 8. In this figure, the third sending node is sending between 70 and 170 s. As can be seen, the sending rates of the two other connections vary more in the middle phase of the test when there are three connections and two bottleneck links in the active state.

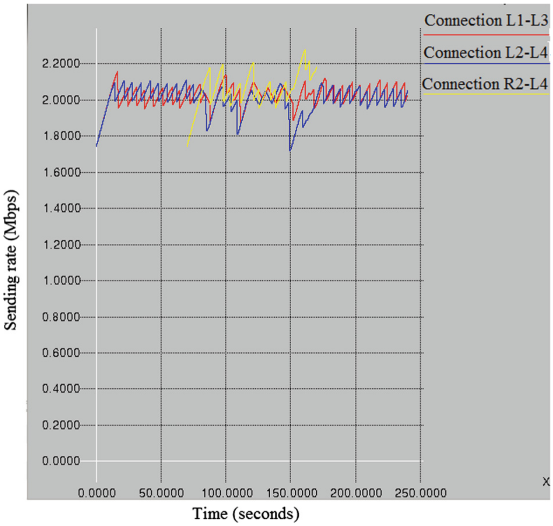


Fig. 8. Test case considering two bottleneck links.

4.6 Advantages of the Minimum Delay Value Pushing

Delay-based congestion control mechanisms have some well-known problems which can affect their performance. The papers [25, 26] list and analyze these problems. Based on these papers, the common problems of delay-based congestion control mechanisms are:

- inability to compete fairly against loss-based congestion control protocols
- persistent congestion
- clock synchronization problem if one-way delay measurements are used
- rerouting problem.

In this subsection, CVIHIS' capability to cope with these problems is explained, although CVIHIS' real-time mode is not a pure delay-based mechanism. It is important to note that these problems can be solved with the help of the minimum delay value pushing in CVIHIS.

Competing against loss-based congestion control protocols is not a big problem for the real-time mode of CVIHIS. Here, CVIHIS was tested against the loss-based TCP protocol. It was observed that the real-time mode of CVIHIS is actually capable in competing against TCP. This mode can compete against loss-based algorithms because CVIHIS shifts its target delay area upwards when competing behavior is needed.

In persistent congestion, the queue of the router is occupied all the time. As a result, delay-based congestion control mechanisms cannot obtain proper value of the minimum delay. The paper [25] suggest that shifting of the minimum delay value alleviates the persistent congestion problem. CVIHIS pushes its delay areas upwards by shifting the minimum delay value. This increases the congestion level of the network leading finally to packet drops. Many connections back off after these packet drops. They reduce their sending rates in a multiplicative manner and the congestion level of the network alleviates. This allows CVIHIS to estimate the correct value of the minimum delay.

The problem with measuring one-way delays is that the clocks of the devices are typically not synchronized accurately in the Internet. Therefore, the one-way delay measurement includes the corresponding one-way delay and the clock offset between the nodes. Even if initially accurately synchronized, two clocks will differ after some time due to clock drift. Due to clock offset and clock drift, one-way delay measurements are challenging.

For CVIHIS, clock offset is not a problem as CVIHIS probes two delay values, minDelay and maxDelay (see Fig. 1) and divides the delay space between these values into several delay areas. CVIHIS can do this correctly if maxDelay is greater than minDelay even if these delay values are negative due to the clock offset. The actual one-way delay measurement related to a certain packet includes the same clock offset and, therefore, the calculated delay is within the minDelay-maxDelay area.

Clock drift, however, can cause problems for CVIHIS. If the measured delay value is increasing due to clock drift, the minDelay value will become outdated. After a certain period of time, the minDelay value does not correspond to the actual propagation delay of the connection path any more. In an extreme situation, the measured delay value including only the propagation delay component may reside closer to maxDelay than minDelay. This means that the connection makes a conclusion that there is an incipient congestion in the network although the queues of the routers are completely empty. This problem can be solved by updating the minDelay value from time to time. In this way, pushing the minimum delay value upwards helps to cope with the clock drift problem. The maxDelay value will be updated after every packet drop, therefore, the clock drift is not critical for the maxDelay. The real network tests of CVIHIS indicated that the clock synchronization problem is not harmful for CVIHIS because the tests were carried out without the synchronization of clocks.

If the route of a connection is changed without an explicit signal from the network, the end host cannot detect it. If the new route has a shorter propagation delay, this does not cause any serious problem for CVIHIS as some packets will probably experience shorter one-way delay values and the minimum delay value will be updated. The maximum delay value will also be updated after the next packet drop. On the other hand, if the new route has a longer propagation delay than the original one, it can pose a problem to CVIHIS. The connection cannot know whether the increase in the delay is due to a congestion in the network or change of the route. Without this knowledge, the end host will interpret the increased delay as a signal of a congestion and the host will decrease its sending rate.

In the following, the rerouting properties of CVIHIS are tested using a simulation. This test uses the real-time mode version of CVIHIS. The ability of CVIHIS to discover rerouting is based on pushing of the minimum delay value and updating the maximum delay value. The maximum delay value is updated after every packet drop. In the simulation, there are two possible routes between the end nodes. There is a direct default route, which is switched off two times during the simulation. So, the traffic has to be switched to the backup route, which has longer propagation delay than the default route. The default route is switched off between seconds 80–150 and 220–320. The capacity of the default route is 700 kbps and the capacity of the backup route is 400 kbps. The simulation result is presented in Fig. 9. As can be seen, CVIHIS can observe the route changes and it can accommodate its sending rate according to the new route.

The problems of delay-based congestion control mechanisms suggest that it is perhaps useful to update the minimum delay value from time to time also in the backward loading mode. Continuous shifting of the minimum delay value could be used also with the backward loading mode. Of course, the shifting factor of this mode should be only slightly over one as we do not want to compete against other connections with this mode. The paper [25] introduce another kind of solution for updating the minimum delay value. After receiving a certain number of data packets, the receiver can check the smallest delay value among these packets. If the difference between the smallest delay value and the current minimum delay value is larger than a certain threshold for a certain number of consecutive times, the receiver interprets this as a change of the propagation delay.

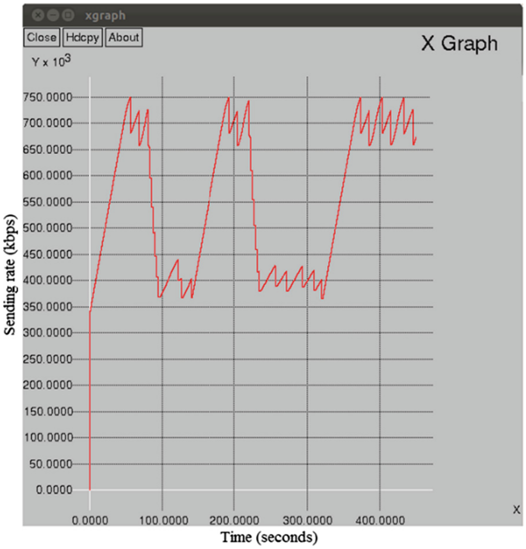


Fig. 9. Rerouting test of the real-time mode.

5 Conclusions

During the last decade, video-type data services in their various forms have become increasingly common. In certain parts of the network, this type of data transmission generates considerably more than half of the total network traffic. We have developed a congestion control mechanism, which is particularly suitable for long-living video transfer. This mechanism includes two modes, the backward loading mode and the real-time mode.

The main objective of the backward loading mode is to back off when there are bandwidth demands from other connections. Based on the test cases, we can conclude that the backward loading mode operates primarily as expected. This mode gives bandwidth away to other connections when the load level of the network is high enough. The main objective of the real-time mode is that it should be TCP-friendly. At the same time, however, it is desirable that the sending rate of this mode would vary in a much smoother way than TCP's sending rate. Based on the tests, it can be said that these objectives are met, however, as usually with this kind of solutions, not in a perfect way. The developed mechanism could manage better regarding TCP friendliness when short or long round-trip times are considered. On the other hand, it is a deliberate decision to change the sending rate of CVIHIS based on the square root of the round-trip time instead of using TCP's round-trip time approach.

The current state of the Internet presents challenges related to the proper operation of Internet congestion control. The Internet is a very heterogeneous environment with different types of applications sending different amounts of data through different kinds of network paths. Based on our research results, relatively small differences between the TCP versions can challenge TCP friendliness. For this kind of heterogeneous environment, the only well-functioning congestion control solution seems to be the currently widely used network overprovisioning. Based on our research we suggest that Internet congestion control can be put into practice in a well-functioning way if there are only a few compatible congestion control mechanisms present. Thus, it is important that a congestion control mechanism behaves against itself in a fair manner. The results of this study show that the CVIHIS connections are able to share the network capacity with each other in a fair manner.

Several research directions can be mentioned for future work. In all the test cases of this study, fixed packet size was used. There is a need to take into consideration heterogenous packet sizes. For example, rate adaptation could be based on the number of bytes rather than the number of packets. There is also a need for research collaboration with specialists of other research areas. One such kind of problem is how the rate adaptation behavior of the application could be integrated with the mechanism of CVIHIS. It is an application specific issue if the real-time mode is smooth enough for the needs of a particular application, because the sending rate of this mode oscillates slightly due to the pushing of the minimum delay value.

References

1. Cisco: Cisco Visual Networking Index: Forecast and Methodology, 2015–2020 (2016). http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html. Accessed 02 Feb 2016
2. Donkor, F.: The comparative instructional effectiveness of print-based instructional and video-based materials for teaching practical skills at a distance. Int. Rev. Res. Open Distance Learn. **11**(1), 96–115 (2010)
3. Pandey, A., Patni, N., Singh, M., Sood, A., Singhd, G.: YouTube as a source of information on the H1N1 influenza pandemic. Am. J. Prev. Med. **38**(3), 1–3 (2010)
4. Bianzino, M., Chaudet, C., Rossi, D., Rougier, J.: A survey of green networking research. IEEE Commun. Surv. Tutor. **14**(1), 3–20 (2013)
5. Vihervaara, J., Loula, P.: Dual-mode congestion control mechanism for video service. In: 7th International Conference on Information and Multimedia Technology, ICIMT 2015, pp. 50–56 (2015)
6. Vihervaara, J., Alapaholuoma, T., Loula, P.: Dual-priority congestion control mechanism for video services, real network tests of CVIHIS. In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 3, pp. 51–59. KMIS (2016)
7. Floyd, S., Fall, K.: Promoting the use of end-to-end congestion control in the Internet. IEEE/ACM Trans. Netw. **7**(4), 458–472 (1999)
8. Singh, K., Yadav, R., Manjul, M., Dhir, R.: Bandwidth delay quality parameter based multicast congestion control. In: 16th International Conference on Advanced Computing and Communications, IEEE ADCOM 2008, pp. 399–405 (2008)
9. Kurose, J., Ross, K.: Computer Networking: A Top-Down Approach, 6th edn. Pearson International Edition, London (2012)
10. Lakshmi, G., Bindu, C.: A queuing model for congestion control and reliable data transfer in cable access networks. Int. J. Comput. Sci. Inf. Technol. **2**(4), 1427–1433 (2011)
11. Braden, R., et al.: Recommendations on queue management and congestion avoidance in the Internet. In: IETF RFC 2309, Informational (1998)

12. Shalunov, S., Hazel, G., Iyengar, J., Kuehlewind, M.: Low Extra Delay Background Transport (LEDBAT). IETF RFC6817 (2012). https://tools.ietf.org/html/rfc6817. Accessed 06 June 2016
13. Kuzmanovic, A., Knightly, E.: TCP-LP: low-priority service via end-point congestion control. IEEE/ACM Trans. Netw. **14**, 739–752 (2006)
14. Kohler, E., Handley, M., Floyd, S.: Datagram Congestion Control Protocol (DCCP). IETF RFC4340 (2006). https://www.ietf.org/rfc/rfc4340.txt
15. Floyd, S., Handley, M., Padhye, J., Widmer, J.: TCP Friendly Rate Control (TFRC): protocol specification. IETF RFC5348 (2008). https://www.ietf.org/rfc/rfc5348.txt. Accessed 06 June 2016
16. Holmer, S., Lundin, H., Carlucci, G., De Cicco, L., Mascolo, S.: A Google congestion control algorithm for real-time communication on the World Wide Web. IETF Informational Internet Draft (2015). https://tools.ietf.org/html/draft-alvestrand-rmcat-congestion-03. Accessed 02 Feb 2017
17. Widmer, H., Denda, R., Mauve, M.: A survey on TCP-friendly congestion control. IEEE Netw. **15**(3), 28–37 (2001)
18. Braden, R.: Requirements for Internet Hosts—Communication Layers. IETF RFC 1122 (1989)
19. Akan, Ö.: On the throughput analysis of rate-based and window-based congestion control schemes. Comput. Netw. **44**, 701–711 (2004)
20. Ros, D., Welzl, M.: Less-than-best-effort service: a survey of end-to-end approaches. IEEE Commun. Surv. Tutor. **15**(2), 898–908 (2013)
21. Almes, G., Kalidindi, S., Zekauskas, M.: A One-way Delay Metric for IPPM. IETF RFC 2679 (1999)
22. Meddeb, A.: Internet QoS: pieces of the puzzle. IEEE Commun. Mag. **48**(2), 86–94 (2010)
23. tc: tc(8)—Linux manual page (2016). http://man7.org/linux/man-pages/man8/tc.8.html. Accessed 06 June 2016
24. Ha, S., Rhee, I., Xu, L.: CUBIC: a new TCP-friendly high-speed TCP variant. ACM SIGOPS Oper. Syst. Rev. **42**, 64–74 (2008)
25. La, R., Walrand, J., Anantharam, V.: Issues in TCP Vegas (1999). http://www.eecs.berkeley.edu/~ananth/1999-2001/Richard/IssuesInTCPVegas.pdf. Accessed 02 Mar 2017
26. Rodríguez-Pérez, M., Herrería-Alonso, S., Fernández-Veiga, M., López-García, C.: Common problems in delay-based congestion control algorithms: a gallery of solutions. Eur. Trans. Telecommun. **22**(4), 168–178 (2011)

Reframing Coordination in Knowledge Transfer: A Sociomaterial Perspective

Néstor A. Nova(✉) and Rafael A. González(✉)

Department of Systems Engineering, Pontificia Universidad Javeriana,
Bogotá, Colombia
{novanestor, ragonzalez}@javeriana.edu.co

Abstract. Coordination issues in knowledge transfer have become increasingly complex due to technological diversity. A classic view on coordination, based on structured and shared cognitive models, is being replaced by an understanding of how coordination actually happens in practice, in a context-dependent and non-routine way. This paper presents a sociomaterial perspective of both knowledge transfer and coordination. Specifically, through a case study in the heritage domain, we investigated how a sociomaterial perspective drives coordination in inter-organizational knowledge transfer activities. Using an affordance-centered analysis, we explore how the process of imbrication has resulted in ontological transformations in sociomaterial heritage co-evolving with dynamic and emergent coordination practices of the associated knowledge transfer.

Keywords: Coordination · Knowledge transfer · Sociomateriality
Affordances · Imbrication

1 Introduction

This paper is an extended version of work published in [1]. We extend our previous work by proposing sociomateriality as an appropriate way to view and analyze coordination in knowledge transfer. In recent years there have been many advances in knowledge transfer technologies, Web 2.0 and 3.0, cloud computing, mobile communications, among others which have contributed to more accurate and real-time information processing and communication for effective knowledge management. As a result, nowadays, many of the challenges faced in knowledge transfer activities are not caused by a lack of technological resources for coordinating but rather by a diversified use of technology within or between organizations. This situation tests the limits of the dominant information processing view of coordination. In this view, coordination mechanisms are considered stable entities as organizational theorists argue [2–4], whereas in practice they behave as dynamic intermediaries between actors in a fluid sociomaterial environment. “In other words, they shared the belief that coordination can be designed in advance in models shared across an organization and that no improvised acts are needed because the environment is predictable” [5]. In contrast, much prior work in knowledge management assumes a common set of coordination mechanisms as tools and technologies that groups or teams use together in the same

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 346–362, 2019.
https://doi.org/10.1007/978-3-319-99701-8_17

way (e.g. groupware), however empirical works show that people are likely to use different technologies with different members in different ways and at different times, and a tendency to use simpler and familiar coordination tools for sharing information.

Besides this, people interact with different coordination mechanisms according to particular circumstances in order to adapt to conditions of uncertainty, novelty, and change and this behavior shifts the thoughts about how coordination take places in practice where coordinating mechanisms do not arise prior to coordinating but are constituted through coordinating [6]. Recent research in organizational coordination has stressed the importance of shifting the focus from structured and shared cognitive models (e.g. standards, rules, schedules) to how coordination actually happens in practice [7]. In fact, the central concern is how people use circumstances to accomplish intelligent actions, not how they apply shared cognitive models to particular situations [8]. Hence, in settings where work is contextualized and non-routine, traditional models of coordination are insufficient to explain coordination as it occurs in practice [9].

This is particularly true for activities in the architectural heritage domain, where the work setting is constantly changing both temporally and spatially due to the unique features of each heritage object. In this sense, the material heritage object not only has value for its physical, material and tangible dimension but also by the interpretation and understanding from heritage experts, residents and authorities about those dimensions. It follows that the semiotic, contextual and physical features are not independent dimensions but mutually related and so a heritage object enacts a clear example of a sociomaterial reality based on a relational ontology. Each dimension is performed through the emergent practices of knowledge transfer in ad-hoc heterogeneous and collaborative teams, where the epistemic and ontological distance between multidisciplinary specialists magnifies knowledge differences and sets up a barrier for the knowledge transfer process.

This move implies that KM researchers can no longer make an ontological distinction between coordination technologies (or mechanisms more generally) and people so that an alternative coordination approach is needed. The present research is grounded on the proposition that the broad banner of sociomateriality [10–13] presents us with an opportunity for reconceptualize coordination- from thinking about how coordination technologies as discrete artifacts influence the knowledge transfer process between people to examining how coordination mechanisms and actions are materially constituted in the knowledge transfer practice, and they are, thus, sociomaterial in nature.

Although the relation between knowledge transfer and coordination is widely recognized, it has not been studied under the sociomaterial umbrella. Current research on sociomateriality explains how the processes of sociomaterial imbrications [10] or entanglements [13] takes place in diverse environments and organizational settings, especially at intra-organizational level [14], but there is a need for knowing about how sociomaterial practices take place in inter-organizational knowledge transfer projects. Therefore, the research question addressed in this paper is the following: How can a sociomaterial perspective drive coordination in inter-organizational knowledge transfer activities in the heritage domain?

The remainder of this paper is structured as follows. The next section deals with the related theoretical grounding and introduces concepts about coordination. Based on the earlier version of this paper, a set of coordination issues in knowledge transfer activities is displayed. Then, we propose some reasons why coordination should be reframed. Next, the sociomaterial lens to reframe coordination is discussed, focusing in the imbrication metaphor and using an affordance perspective. Afterward, using the sociomaterial perspective, we use three examples of imbrications in the architectural heritage domain and then we posit four propositions about the sociomaterial view of coordination. Finally, some conclusions are drawn acknowledging the limitations and suggesting future research.

2 Coordination Theory

Early theories of coordination focused on the need to balance differentiation among organizational units, with integration being achieved through coordination mechanisms [2–4]. The information-processing -IP paradigm has been a common approach to address coordination in prior research. From this focus, coordination is about the integration of organizational work under conditions of task interdependence and uncertainty [9]. Specifically, coordination is defined as managing dependencies between activities [15]. Interdependencies refer to goal-relevant relationships between activities [16] while coordination is made operational through a set of coordination mechanisms which stipulates and mediates the articulation of activities and provides affordances and constraints to articulation work [17]. All coordination processes include actors performing interdependent activities and interdependencies generate incremental IP needs, but when interdependency is higher, a coordination mechanism can facilitate or affect the IP capability of the organization [1].

According to the IP view, each coordination mechanism represents a specific IP capability and must be matched to the IP needs of the context generated by the interdependence between work units. Thus, the organization can get a prescription to deal with uncertainty and organizational contingencies through a set of mechanisms that can provide higher IP capacity or lower amount of information to be processed. As a result of this perspective, literature has emphasized the distinction between different modes of coordination, for example, by program or feedback [3], impersonal versus mutual adjustment [18], rules or programs, hierarchies and target or goals [2], and programmed versus non-programmed [19], rules and directives, sequencing, routines and group problem solving and decision making [20], task-task, task-resource and resource-resource [21] and structural and formal, hybrid/overlaying, informal and internal markets [22].

More recently, Malone and Crowston developed a coordination theory based on the process level which strength is its recognition of the complexity of interdependencies in organizational work [15, 16]. The alignment between dependencies and mechanisms as a handbook of coordination could provide a guide for analyzing the coordination needs in particular situations and generating alternative ways of fulfilling them [23]. Following the IP view, interdependencies are classified as resource flow, fit and sharing. Flow dependencies occur when an activity produces a resource that is used by other activity. For instance, in cultural architectural heritage activities, task complexity depends on many rules and procedures for managing information such as get approvals from the local or national authorities for a special plan for protecting and managing heritage objects before to start the physical restoration activities. Fit dependencies arise when multiple activities produce a single resource, as an example, let´s consider an special plan for protecting and managing heritage objects whose develop requires the work of architects as experts in the patrimonial scope, but also from other complementary disciplines such as civil engineers, art restorers, anthropologists, ecologist, lawyers, among others, each one performing specific analysis regarding the heritage object state but the final document must integrate different outcomes from each approach [1]. Finally, sharing dependencies occur when multiple activities use the same resource. Following the example above, the final draft of the special plan for protecting and managing heritage objects must be reviewed by the consulting group, local authorities, the community who take care and lives in the heritage object, the object owner, and the funding institution. Thus, the type of interdependence determines the mode of coordination used.

In the IP view, coordination mechanisms can be classified as: standards, mediation and mutual adjustment [2–4] and existing research on coordination has revealed a large number of coordination mechanisms through which coordination actions take place in inter-organizational knowledge transfer projects. Standard-based mechanisms are considered an a priori specification of codified guidelines, action programs and specific goals [3, 4] where the verbal communication and the interaction among actors is not necessary for coordination [2]. Examples in the architectural heritage domain include governmental policies, interventions plans, management plans for protection, thesauri, among others.

Mediation-based mechanisms involve a third actor typically located at a higher level that acts as mediator between two organizational units [24]. Some examples in the architectural heritage domain are technical committees, technical reports, hierarchies, labor division by discipline, GIS, cloud computing, and others. Mutual adjustment mechanisms are based on the expected reciprocal communication between actors [4]. Unlike standards-based mechanisms, communication and interaction in architectural heritage projects is achieved through personal channels between peers such as scheduled and unscheduled meetings, instant messaging, video conference, email, others.

The proposal of Malone et al. [23] shares with the IP view the assumption that the environment is predictable enough to characterize existing interdependencies but predefined mechanisms can be designed for various contingencies [9]. Indeed, coordination mechanisms need to have sufficient flexibility to cope with uncertainty, novelty, complexity and ambiguity. Uncertainly is defined as the difference between the amount of information required to perform the task and the amount of information already possessed by the organization [2]; novelty is featured by the newness of a process [25]; complexity means the number of interrelated elements or sub-systems within the systems and the interdependency between them [4]; and ambiguity refers to lack of understanding between actors [26] during organizational activities and the outputs that they are intended to organize.

3 Coordination Issues in Practice

3.1 Research Context

This study was performed in the architectural heritage domain. Specifically, we have developed a deep exploration of the Iberoamerican Historical Heritage Network – RedPHI, which was constituted in 2011 by seven universities working in material architectural heritage management. RedPHI aims for protection and conservation of the cultural heritage through research and consultancy projects as well as training and divulgation of knowledge. The main objective is to manage and develop conservationist projects in the architectural, urban and landscape scope.

RedPHI Network has some particular characteristics which make it a special setting for studying coordination in the interorganizational context. First of all, the RedPHI work approach demands a multidisciplinary treatment, because is not only explored from its materiality but also its relation with the human being. From the RedPHI scope, three dimensions can be distinguished: semiotic, contextual and physical. The semiotic dimension entails the personal interpretation and appreciation for who studies, protects, maintains and preserves the heritage object and in some cases, for those who use it as their home. The contextual dimension reflects the value of the heritage object, abstracted from its context, analyzed jointly with their environment and observed as a part of a territory. The physical dimension represents their physic, material and tangible being. Thus, not only the objective content but a subjective attitude and a function of the understanding enact the features that distinguish the heritage object from a territorial approach. Following this perspective, the set of dimensions are not independent but mutually related and so a heritage object can be understood as sociomaterial reality.

Second, RedPHI projects are naturally collaborative and inter-organizational in the sense that they involve heterogeneous actors that develop specialized activities. Activities can be performed by multiple disciplines like architects, civil and chemical engineers, anthropologist, biologist, lawyers, among others, which work in different organizational units as universities, research groups, government and private institutions, national and international organizations and networks, and society in general.

Third, among the heterogeneous actors flow different types of information, which change according to: the project phase being developed, the state of the project or activity, the requirements of the client, the findings, and the types of actors that collaborate, among others. Information of activities in heritage projects is associated with: task records, management, standards compliance, formalization and execution of collaborative work and dissemination of knowledge. This information is available mostly in digital documents, only the architectural, structural and electrical blueprints have a strictly material form. Most of the documentary information corresponds to unstructured narrative text and in some cases semi-structured.

3.2 Coordination Issues in Architectural Heritage Activities

Different coordination issues can be identified regarding the role of coordination in knowledge transfer. For example, coordination needs increase when knowledge linked to the cultural heritage is highly complex and ambiguous due to the interaction of knowledge specialized and the strategy of division of labor for performing specific research activities on material cultural heritage. Ambiguity and labor division by disciplines causes a lack of common understanding between the leader and the experts, or even among the experts themselves, which is also a problem for transferring knowledge. Likewise, ambiguity increases because most of the knowledge involved in the domain of material cultural heritage remains hidden, because of lacking of codification and verbalization. In this sense, coordination performance depends on the actor's alignment level when choosing the mechanisms to coordinate, if this alignment is not sufficient, coordination performance depends on the leader's accountability.

In addition, the coordination process follows a non-analytic process addressed by experts' experience and observation capabilities and it does not include any analysis regarding the election effects on the task performance. Moreover, the decision-making about selecting coordination mechanisms follows an experience-based process, where the previous use of the mechanisms is the main argument for selection. For instance, when research is carried out in a heritage domain project, it is assumed that the set of coordination mechanisms used in the last project will be used in the new one, so they do not take into account the situation and contextual features of the new project as requirements for modifying the coordination mechanisms.

Often, coordination mechanisms chosen for managing activities are not always cognitively feasible or applicable for handling different interdependencies. For example, face-to-face meetings is the favorite mechanism for coordinating most of the interdependencies, but it becomes problematic when people are not located inside the university or when there are many people participating in the same activity. In this sense, it is necessary to use Skype in particular for interaction, but sometimes using this technological platform is considered as specialized knowledge and people often rejects its use. Then, people opt for face-to-face meetings between small groups of experts, which make difficult to transfer knowledge among all actors and affects collaboration.

Furthermore, despite the fact that heritage activities are dynamic, coordination mechanisms do not change as fast as. For example, obtaining requirements in a restoration project requires high verbal communication to understand the object and its characteristics and relationship with the environment, and for identifying the customer needs and specifications around the design. The initial stage of a restoration project includes a diagnosis of the current state of the heritage object, therefore a wide set of coordination mechanisms such as technical visits, meetings, discussions, socializations, brainstorming, client meetings and expert meetings are employed. However, during the restoration activities it is necessary to use mechanisms such as technical committees, hierarchies, authority, work plans, instant messaging, telephone calls, among others, in order to contrast periodically the initial architectural specifications with respect to the current state of the process, given the small tolerable margin of error in objects which are of outstanding universal value for humanity.

Although using the same set of coordination mechanisms for all projects is not a problem in itself, it could be problematic when the omission of alternative mechanisms affects task performance. For example, when ontological differences emerge in a research or consulting activity, conventional meetings is the first coordination mechanism applied. If there is no consensus, the project leader divides the project operation into functional subgroups usually by disciplinary area, for example the anthropologists team and the architects team. However, if the differences persist, the leader determines the ontological principles to be followed according to the client's requirements. However, actors are not willing to try other mechanisms that could facilitate common understanding, such as thesauri or glossaries of legal terms, because it is considered a superficial way which limits the expert´s interpretation capabilities, and does not allow adding additional value from contextual analysis and interpretation.

Also, selecting coordination mechanisms is a process driven by the lower cost, this means that the inclusion of new coordination mechanism in the current portfolio is omitted if this involves additional resources allocation. An example of this is when the expert team has previously performed a particular activity which is required in another project. This experience drives experts to assume certain aspects of the heritage object that could reduce coordination costs in obtaining client´s requirements, and also motivates them to use the set of related mechanisms that they have previously used. But if problems arise later in activities and/or results, coordination needs increase, which either reinforces the use of the available mechanisms or forces the inclusion of new ones to correct errors.

Another problem involves the limited combination of the coordination mechanisms available. In a standard and natural way each dependency is coordinated through a set of coordination mechanisms that are difficult to combine with others of the portfolio and thus it can affect task performance. For instance, mechanisms such as hierarchies, authority and working documents are used for decision making during the planning stage, but it could also include mechanisms for project assessment, expert communities, discussion and debates, knowledge bases, KMS, among others.

Additionally, the specification of coordination mechanisms should be proportional to the dependencies to be managed, but sometimes the specification is not appropriate. An example of over-specification was evidenced in multidisciplinary activities which involve distinct working methods. This dependency was attempted to coordinate through a Wiki. However, the Wiki became an additional database and it did not enable collaboration, because the platform was unknown for most of the experts, so the initial purpose of mutual adjustment for decision-making was reduced to a complementary work support activity. An example of a sub-specification was observed in a Master's program in cultural heritage in one of the RedPHI members. The Masters approach was modified in order to update the ontological and epistemological focus of heritage from monument vision to the territory perspective, and this changes involved a deeper discussion among teachers, but they did not take into account the students, graduates or hiring companies' opinions, therefore the scope of the coordination mechanism was not adequate and its effect on performance was not what was expected.

Lastly, familiarity, availability, confidence, experience, natural and routine use, upgrading facilities are criteria for selecting a set of technologies that support coordination activities. Due to the overload of ICTs for supporting coordination, each actor uses different technologies, in different ways, at different times and with different people, so this shows that selecting tools is a highly situated and contextualized process. In addition, this behavior affects the mutual understanding between actors which in turn, affects knowledge transfer. For addressing this divergence in ICT usage, experts with prior experience collaborating among themselves, state a common denominator of tools and new actors must adjust to the preexisting ICT tools.

Nevertheless, the common denominator can be too limited and eventually insufficient to transfer the amount of knowledge and information that the project requires. In addition, actors minimize information exchange to avoid learning new tools, due to resource availability.

4 Request for Reframing Coordination

The above uncovers a gap between the conception and use of ICTs that support coordination, and a lack of understanding about how this gap affects the knowledge transfer process. This issue exceeds the scope of the mainstream IP view of coordination, because coordination problems in knowledge transfer are not a matter of information quantity or IP capacity, but a relationship between people and coordination technologies. This point overcomes the techno-centric view of coordination, which has been widely studied, and suggests the need to reframe coordination.

In architectural heritage projects, experts use different coordination tools and technologies for knowledge transfer with different members of their research or consultancy project and often selecting coordination mechanisms depends on the context, project type, prior relationships, familiarity, cost, among other reasons and frequently the coordination process supports dyadic relations between people in the same discipline, faculty, university, or research group. Accordingly, selecting coordination mechanisms is a dynamic and volatile process difficult to convert into coordination patterns. Coordination practices based on a handbook of coordination mechanisms [23], seems less appropriate for knowledge transfer because people are likely to use different technologies with different members in different ways and at different times [27], using multiple technologies simultaneously and switching the technologies they use frequently [28] and a given technology is often used interdependently with a wide swath of other pervasive and embedded technologies [29]. According to DeSanctis and Poole, people within the same social group (as opposed to across social groups like interpretive field studies normally show) sometimes used the same technology in ways that were different from their group members [30]. In addition, groups interpret technologies in different ways, based on the social contexts in which they encounter them [31], for instance, people in two different organizations use the same new technology differently and, consequently, change (or not) their informal organizing in distinct ways [32].

A human agency perspective suggests that people's work is not determined by the technologies they employ [10]. As Orlikowski notes, people "have the option, at any moment and within existing conditions and materials, to 'choose to do otherwise' with the technology at hand" [33]. Nowadays, the social web, ubiquitous computing, haptic devices and digital artifacts become increasingly interactive, reprogrammable, editable, distributed, mobilizable across borders and settings [34]. This properties are enacted through the recurrent use of a technology. They are not embodied within the technology; rather, they emerge from the ongoing and situated interactions that users have with the technology at hand.

In addition, technology is able to perform tasks in ways which are transparent and frequently unnoticed by humans [10]. As a result of this new perspective (and the inherent properties), artifacts can be invoked in different ways, take initiatives (sometimes without explicit human intervention, but in response to other digital artifacts), act upon ‘digital matter’ or even create ‘digital matter’ for others to account for [35]. This suggest examining technology use at the level of the feature instead of at the level of the artifact, because there are considerable variations in what features people perceive and use or what technology affords them. We respond to the preceding aspects by using a sociomaterial approach to shift attention to the question of how coordination mechanisms enact an active member of the social network that comprises the knowledge transfer process and it proposes a reorientation of knowledge coordination away from pre-identified interdependences and coordination mechanisms. This reframing is necessary and timely because of the growing recognition that knowledge is inextricably bound up with the material and social context in which it is transferred and acquired [8].

5 Sociomateriality

The theoretical perspective of sociomateriality results from various attempts to challenge the separation between humans and technologies, and it enriches our understanding of their interplay and dynamics. This duality perspective has dominated IS research but it presents a conceptual difficulty when faced with the increasingly complex materiality of everyday IS-mediated work practices [12, 36]. Sociomateriality can be considered as an “analytical break” that can help us avoid the dichotomy that exists between the social and the technical [28]. The concept of sociomateriality in the information systems discipline rests on the seminal work of Orlikowski and Scott [12, 13, 37] and Leonardi [10, 11, 32] and it draws on the work of actor-network theory - ANT [38–40].

The conception of sociomateriality “makes a distinctive move away from seeing actors and objects as primarily self-contained entities that influence each other…away from discrete entities of people and technology…to composite and shifting assemblages” [13]. While a separateness ontology assumes that human beings and things—the social and the material—exist as separate and self-contained entities that interact and affect each other, a relational ontology assumes that the social and the material are inherently inseparable [13]. In this line, entities, human beings, and things exist only in relations: they are performed and continuously brought into being through social relations [37, 39].

Based on different ontologies, scholars are just beginning to discuss how the relation between agencies occurs in practice. Literature discusses the relation between agencies as entanglement [12], assemblage [41], or imbrication [10]. The metaphor of imbrication describes the arrangement of distinct elements in overlapping patterns so that they function interdependently. Ontologically, imbrication assumes that components of a sociomaterial assemblage can be disentangled, separately improved and then re-arranged, but it depends on human agency to perceive affordances and constrains that make people unfold diverse intentions for using technology. People have the ability to perceive whether an artifact offers diverse possibilities for action or limits their ability to carry out their goals. According to Leonardi [10], imbrication occurs in two ways, first, when an affordance is perceived from a technological artifact an existing material agency is imbricated with a new human agency and people may be likely to change their routine. Second, when a constraint is perceived form a technological artifact, an existing human agency is imbricated with a new material agency and then a technology changes.

Affordances are rooted in a relational ontology which gives equal play to the material as well as the social [42]. Affordances are by definition a sociomaterial construct and therefore studying affordances—through untangling the complex interactions between multiple social actors and material artifacts—is one potential approach to empirically analyzing sociomateriality [36]. The concept of affordances, refers to the action potential that can be taken given a technology [10, 43, 44]. Several researchers have advocated this concept as a promising approach to overcome the subject–object and agency–structure oppositions that have restrained much of the research at the intersection of technology and organizations [10, 45, 46]. By treating the entanglement between the human action and the technological capability as a unit of analysis, the affordance perspective provides a language for beginning to examine coordination activities and its role in affecting the process of knowledge transferring [47]. Our use of the term affordance is in line with Leonardi [10] focusing on how people draw on infrastructure to construct a perception that a technology either constrains their ability to achieve their goals, or that the technology affords the possibility of achieving new goals. For understanding the affordances theory in the domain of architectural heritage work, next we present an example of affordances in practice.

5.1 An Empirical Illustration: Performing an Architectural Survey in Architectural Heritage Objects

Imbrication 1 – Hose-Based Level to Total Station. For many years, the hose-based level has been the common way to measure settlement displacements with a high level of accuracy using a simple tool. Two communicating vessels are interlinked with each other by means of a hose, which contains a liquid. The level of the liquid is read off by analogue means on a scale. Although the system is simple to use and can be easily adapted for different areas, some mistakes could arise due to water leaks or wrong water levels. Thus, the hose-based level tools has a margin of error that can affect leveling studies in architectural heritage projects. The constraint explained above was leaded by the experts changing the hose-based level by a total station, which is a surveying equipment combination of electromagnetic distance measuring instrument and electronic theodolite. Experts interpreted total station ability to measure level digitally as affording them to reduce manual errors involved in reading and recording which affect considerably the technical process performed in a heritage project. Then, most of the architects and civil engineers started to use a total station tool in all architectural heritage projects in order to calculate with more accuracy the land level in a heritage object. The new work routine made the fieldwork activities less manual and encourage experts to take advantage of a new technology for improving quality in their heritage studies.

The use of the total station tool changed the work routine but at the same time generated new configurations to coordinate the knowledge transfer. Based on the real time information produced by the team of four architects about design, measures, materials and state, data collected manually from the hose-based leveling was mostly sketched by the architect leader in paper and then the team got feedback with results. The set of architectural drawings enacted a coordination mechanism through which architects shared knowledge each other. The coordination mechanism was used to connect experience of architects and civil engineers with knowledge from other disciplines to complement or expand the heritage analysis. Based on the architectural drawings, sometimes it was necessary to hire additional experts i.e. topographer, so as to verify the field measures and confirm the survey results. Later, with the final version of the architectural drawing, the architects built a mockup (a 3D model elaborated manually) representing the current state of the object which afforded them to make decisions about the task sequence for the next stage project, usually the restoration activities. So making decisions during the whole process took a lot of time. Finally, the change in technology enacted another way to exchange information and transfer knowledge because data collected form the total station could be downloaded to computers for further processing and then it was possible to send file outputs by email or share it using cloud computing. Thus, using technological devices afford experts to share new findings easier than citing people in technical committees.

Imbrication 2 – Total Station to 3D Scanner. Although total station allows experts to carry out activities faster than using the hose-based level, a new constraint was perceived by workers. Using total station does not permit to optimize the execution time required to measure by hand all the details that a specific study needs to represent. Thus, the architectural survey took a long time making the study inefficiently, whereby engineers had to rent additional tools affecting the budget of the project. It became clear to experts that total station constrained their ability to perform activities at the scheduled time. Consequently, architects and civil engineers stopped using total station and started to use 3D Scanner. Scanners automatically acquire a so-called ‘point cloud’ with a resolution determined by the surveyor before starting the scan, the clouds acquired will then be connected to each other through the overlapping areas acquired from several different station points, using software that perform this type of calculation that is defined in jargon ‘cloud-to-cloud registration’. Accordingly, when experts began to use this new technology they found that the time spent for data capture in an architectural survey was reduce significantly and the diagnosis and restoration time stop was shorter. As an example, using a 3D scanner in the “Voto Nacional Church”[1] enable experts to detect in one day a deviation in the position of the main religious images on the top of the dome, but using other techniques this diagnostic would had taken some weeks. Also, 3D scanners affords experts the ability to make deeper analysis of the object measures and there is no need to touch the measured object, which in some cases is not possible or desirable. Nowadays, experts use 3D scanners depending on the heritage objects complexity (access, physic state, availability) and as a way to save time.

[1] The Basilica of the Sacred Heart of Jesus, also known as the “Voto Nacional” Church is a Catholic basilica located in Bogotá, Colombia. The church was completed in 1916 and it was declared as a national monument in 1975.

Using total station afford the experts to change the method to collect data from manual to digital techniques, enabling them the use the CAD tools as AutoCAD. Now, the CAD design of the heritage object enact another coordination mechanism. A set of iterations for designing are necessary to get the digital architectural drawing of a heritage object. Now it is necessary to exchange information not only between architects, but also with the CAD expert for preparing digital layouts and sketches and determine design elements for various structures and building components. The skill of drawing has been lost by the architect and now is assumed by the digital design specialist who is considered as another expert. In the technical committees, people engage easier viewing the 3D model than the architectural drawing in paper, and urgent decisions can be made. Using the 3D scanner, the need for drawing in AutoCAD is reduced because the model is already built in less than thirty minutes, and using plugins the software is able to synchronize file outputs with others applications and devices like 3D printers, thus, architects no longer need to build mockups to make decisions because the 3D printer build it faster. Based on 3D models, experts can prioritize tasks and budgeting from first aid interventions to additional details that can be omitted without affecting the final result, and this allows experts to rescue the heritage object, like a patient in an emergency room. Finally, the change in technology enact another way to make decisions emphasizing the importance of the digital design rather than the physical object.

Imbrication 3a – 3D Scanner to Virtual Reality. 3D scanner affords experts many possibilities in order to capture information from complex objects, however 3D scanners do not afford experts to get information from inside of walls, roofs, support beams and others materials of the structure which is very important to detect water, drainage or electricity networks as well as gas connections. Experts perceive this as constraint because they had to use additional tools like radar detectors or x-ray scanners which afford them to explore internal conditions of the object structure making the tests less invasive than 20 years ago when experts must break the wall to perform vulnerability studies. But post processing the point cloud is another constraint, because there is enormous work for manipulating and filtering many points in order to get useable information before transporting this data into a 3D model. Although most of the software platforms have combined algorithms for triangulation and surfacing of the point cloud, manipulating and filtering data require long time as well as specialized knowledge about how to edit the point cloud. Thus, a new change in technologies is being planned by RedPHI experts for next years. By using virtual reality, experts and people in general can perceive the third dimension and the "virtual investigation" of the object become more realistic. Experts argue that this can be a more simple, natural and correct way to analyze and present information reducing the possibility to make evaluation mistakes due to the false prospective of the classic visualization trough hose-based level, total stations and 3D scanners. So virtual reality can afford experts and non-experts visualizing in only one tool all the technical information like drawings, topographic studies, material specifications, pathological studies, etc., but also can offer information to students, teachers, researchers, majors, specialists, entities, enthusiastic people, and others about history, evolution, transformation, representation, projects performed, cultural expressions, and all kind of information linked to territorial studies.

Imbrication 3b – 3D Scanner to Total Station or Hose-Based Level. The 3D scanner affords experts to make more productive their work with high information quality and allowing them the possibility to represent the space detected in an innovative way different to the traditional 2D. File outputs from a scanner can be processed using diverse CAD tools producing complex and well elaborated models. Many technological tools can be used under a license code for University membership which they keep updated in order to teach students how to use this tools in the real life. However, most of the governmental customers (ministry, secretaries and institutes regarding heritage domain) prefers to analyze the traditional file outputs in 2D and physical drawings rather than 3D models because they do not have technological capabilities to process and analyze it. Consequently, governmental entities consider 3D models as complementary information because for many years they have made decisions without it, moreover bidding conditions include mandatory to submit the architectural drawing in paper. Thus, customer rules and governmental limitations constraint the possibility to take advantage of the digital tools for an architectural surveys, but also they delay the implementation of technological advances that affords many action possibilities for supporting world and national heritage declaratory activities. Additionally, this factors limit the proactive efforts of surveying firms with scanning capabilities. As an effect of this constraint, at the time experts want to participate in a project financed by governmental entities, they have to change their work routine from using high-end technology to using a total station tool, in order to accomplish tasks in the terms demanded by the bidding conditions.

6 Discussion

This paper argues that the coordination issues in knowledge transfer can be overcome if sociomaterial analysis is taken into account. Sociomaterial imbrications of technologies for coordinating and work routines can only be fully understood if changing the coordination emphasis from the how (i.e., the mode) of coordination towards the what (content) and when (circumstances) of coordination. Following the imbrications framework [10] we analyzed a case study of RedPHI sociomaterial imbrications and technologies in the variety and dynamics of the field work performed by architectural heritage experts.

As a result, three main phases of RedPHI fieldwork with different patterns of sociomaterial imbrications were distinguished. This section compares and discusses the results obtained through the phases. The illustrative example shows that imbrications 1, 2 and 3a are characterized by coordination, making decisions, organizational structure and culture, capabilities and institutional influences, and generally support the findings by Leonardi [10] about the patterns of sociomaterial imbrications. In particular, the first three phases suggest that perceptions of constraint leads people to change their technologies while perceptions of affordance lead people to change their routines [10]. However, during the imbrication 3b characterized by coordination, capabilities and institutional influences the pattern described above changed: the perception of constraints and affordances mainly led to change in organizational routines and the characteristics of imbrication 1 and 2 emerged again. Despite that coordination and institutional influences were present at all phases the intensity of change in the institutional links and the intensity of change in the perception of affordances and constraints within an architectural survey significantly changed during imbrication 3b. Governmental requirements forced the experts to use coordination mechanism that have been replaced by changing technology in imbrications 1 and 2 and also work routines had to be transformed but this time not for taking advantage of affordances of new technologies but to coordinate work routines as before, using architectural drawings in paper.

This finding shows that technological affordances can be exploited or not according to contexts and situation and so experts must be willing to modify their coordination strategy and mechanisms in order to accomplish tasks following the project specifications, even if experts keep capacities to produce deeper results and technology affords possibilities to do tasks with greater impact. Our propositions are, thus, the following. First, context and situation (i.e. technology availability, institutional rules, and interdisciplinary work) significantly influence the dynamics of sociomaterial imbrications of technologies and routines within and between organizations. For example, the development process of an architectural survey were significantly influenced by the fact that technology was not available for the architects, by a group of experts able to share knowledge with other disciplines in order to complement the tasks and by institutional requirements which lead to change coordination mechanisms. Second, patterns of sociomaterial imbrication (i.e. change in technology or change in routines) depends on the perception of affordances and constraints within the organization but also largely depends on interconnection between sociomaterial networks of actors participating in the same architectural heritage project. When projects involve heterogeneous actors, the imbrication process of human and material agencies is affected by perceptions of affordances and constraints in other organizations with different context, situations, characteristics, people, technology, process, and this makes the sociomaterial coordination process more dynamic. Third, imbrications of human and material agencies constantly produce novel changes in routines and technologies but in some cases, agencies can be disentangled, reconfigured and entangled following previous imbrications and according to particular requirements. In this sense, the rejection of one technology can simultaneously constitute a new change in technology as Leonardi [10] posits, but also a change in routine is possible as well. Thus, the way that the different actors' networks are intertwined determines the scope of new changes in routines and technologies. Fourth, task context and situation lead to changes in routines and technologies and thus alternative imbrications can be developed within a project. Due to the technological diversity available among heterogeneous actors, it is possible to adapt the work routines in multiple forms using the huge portfolio of technologies in different ways. Thus, the sociomaterial imbrication process cannot be considered as a linear and particular sequence [10] but as a dynamic and emergent process led by ubiquitous and pervasive technology available for all actors.

7 Conclusion

In this paper, we use a sociomaterial perspective on the coordination to enrich our understanding of how imbrications of technologies that enable knowledge transfer relationships. Specifically, findings build a link between coordination technologies, knowledge transfer routines and the diversity of contexts and situations where coordination takes place. As this paper has shown, the increasing diversity of coordination technologies and the flexibility of knowledge transfer routines affords an opportunity to look more closely the coordination issues in which human and material agencies change in response to one another and change according to contextual coordination dynamics. In addition, the affordance perspective brings the opportunity to study coordination from the sociomaterial lens, because this approach is not based on information processing but is practice-oriented which is distinctive of dynamic and emergent coordination studies and affords possibilities to overcome the analytical language of separateness between coordination mechanisms and interdependencies by a relational language in the sense of ANT, which suggest that all coordination practices (including knowledge transfer practices) are always configured by some specific sociomateriality, and thus to study coordination in the knowledge transfer process, we must study the dynamic and emergent sociomaterial (re)configurations as coordination activities performed in practice.

Our findings are based on one case study and, therefore, by definition, only meet to a limited extent the criterion of generalizability. Further research needs to be conducted in other domains. In order to avoid over complexity of this analysis we did not provide additional specific details on knowledge transfer relationships from a sociomaterial perspective, but findings reveal a need for a deeper understanding of knowledge transfer practices following an affordances perspective.

References

1. Nova, N.A., Gonzalez, R.A.: Coordination problems in knowledge transfer: a case study of inter-organizational projects. In: 8th International Conference on Knowledge Management and Information Sharing November 23 (2016)
2. Galbraith, J.R.: Organization design: an information processing view. Interfaces **4**, 28–36 (1974)
3. March, J.G., Simon, H.A.: Organizations. Wiley Organizations, Oxford (1958)
4. Thompson, J.D.: Organizations in Action: Social Science Bases of Administrative Theory. Transaction Publishers, Piscataway (1967)
5. Constantinides, P., Barrett, M.: A narrative networks approach to understanding coordination practices in emergency response. Inf. Organ. **22**, 273–294 (2012)
6. Jarzabkowski, P.A., Lê, J.K., Feldman, M.S.: Toward a theory of coordinating: creating coordinating mechanisms in practice. Organ. Sci. **23**, 907–927 (2011)
7. Okhuysen, G.A., Bechky, B.A.: 10 coordination in organizations: an integrative perspective. Acad. Manag. Ann. **3**, 463–502 (2009)
8. Gherardi, S., Nicolini, D.: To transfer is to transform: the circulation of safety knowledge. Organization **7**, 329–348 (2000)

9. Faraj, S., Xiao, Y.: Coordination in fast-response organizations. Manag. Sci. **52**, 1155–1169 (2006)
10. Leonardi, P.M.: When flexible routines meet flexible technologies: affordance, constraint, and the imbrication of human and material agencies. MIS Q. **35**, 147–168 (2011)
11. Leonardi, P.M.: Theoretical foundations for the study of sociomateriality. Inf. Organ. **23**, 59–76 (2013)
12. Orlikowski, W.J.: Sociomaterial practices: exploring technology at work. Organ. Stud. **28**, 1435–1448 (2007)
13. Orlikowski, W.J., Scott, S.V.: Sociomateriality: challenging the separation of technology work and organization. Ann. Acad. Manag. **2**, 433–474 (2008)
14. Zorina, A.P., Avison, D.: When environment matters: inter-organizational effects on sociomaterial imbrications and change (2011)
15. Malone, T.W., Crowston, K.: The interdisciplinary study of coordination. ACM Comput. Surv. CSUR. **26**, 87–119 (1994)
16. Malone, T.W., Crowston, K.: What is coordination theory and how can it help design cooperative work systems? In: Proceedings of the 1990 ACM Conference on Computer-Supported Cooperative Work, pp. 357–370. ACM, New York, NY, USA (1990)
17. Cabitza, F., Simone, C.: Affording mechanisms: an integrated view of coordination and knowledge management. Comput. Support. Coop. Work CSCW **21**, 227–260 (2011)
18. Van de Ven, A.H., Delbecq, A.L., Koenig Jr., R.: Determinants of coordination modes within organizations. Am. Sociol. Rev. **41**, 322–338 (1976)
19. Argote, L.: Input uncertainty and organizational coordination in hospital emergency units. Adm. Sci. Q. **27**, 420–434 (1982)
20. Grant, R.M.: Toward a knowledge-based theory of the firm. Strateg. Manag. J. **17**, 109–122 (1996)
21. Crowston, K.: A taxonomy of organizational dependencies and coordination mechanisms. In: Organizing Business Knowledge: The Mit Process Handbook. Center for Coordination Science, Alfred P. Sloan School of Management, Massachusetts Institute of Technology (1994)
22. Reger, G.: How R&D is coordinated in Japanese and European multinationals. RD Manag. **29**, 71–88 (1999)
23. Malone, T.W., et al.: Tools for inventing organizations: toward a handbook of organizational processes. Manag. Sci. **45**, 425–443 (1999)
24. Gonzalez, R.A.: Coordination and its ICT support in crisis response: confronting the information-processing view of coordination with a case study. In: 41st International Conference on System Sciences, IEEE Xplore (2008)
25. Adler, P.S.: Interdepartmental interdependence and coordination: the case of the design/manufacturing interface. Organ. Sci. **6**, 147–167 (1995)
26. Simonin, B.L.: Ambiguity and the process of knowledge transfer in strategic alliances. Strateg. Manag. J. **20**, 595–623 (1999)
27. Cummings, J.N., Kiesler, S.: Who collaborates successfully? Prior experience reduces collaboration barriers in distributed interdisciplinary research. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, pp. 437–446. ACM (2008)
28. Contractor, N., Monge, P., Leonardi, P.M.: Network theory multidimensional networks and the dynamics of sociomateriality: bringing technology inside the network. Int. J. Commun. **5**, 682–720 (2011)
29. Bailey, D.E., Leonardi, P.M., Chong, J.: Minding the gaps: understanding technology interdependence and coordination in knowledge work. Organ. Sci. **21**, 713–730 (2010)
30. DeSanctis, G., Poole, M.S.: Capturing the complexity in advanced technology use: adaptive structuration theory. Organ. Sci. **5**, 121–147 (1994)

31. Bechky, B.A.: Sharing meaning across occupational communities: the transformation of understanding on a production floor. Organ. Sci. **14**, 312–330 (2003)
32. Leonardi, P.M.: Materiality, sociomateriality, and socio-technical systems: what do these terms mean? How are they related? Do we need them? In: Materiality and Organizing: Social Interaction in a Technological World, pp.25–48 (2012)
33. Orlikowski, W.J.: Using technology and constituting structures: a practice lens for studying technology in organizations. Organ. Sci. **11**, 404–428 (2000)
34. Kallinikos, J., Aaltonen, A., Marton, A.: The ambivalent ontology of digital artifacts. MIS Q. **37**, 357–370 (2013)
35. Ktistakis, G., Akoumianakis, D.: Digital calendars for flexible organizational routines. J. Enterp. Inf. Manag. **30**, 476–502 (2017)
36. Leonardi, P.M., Barley, S.R.: Materiality and change: Challenges to building better theory about technology and organizing. Inf. Organ. **18**, 159–176 (2008)
37. Orlikowski, W.J.: The sociomateriality of organisational life: considering technology in management research. Camb. J. Econ. **34**, 125–141 (2009)
38. Callon, M.: Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. Sociol. Rev. **32**, 196–233 (1984)
39. Latour, B.: Reassembling the Social: An Introduction to Actor-Network-Theory. OUP, Oxford (2005)
40. Law, J.: Technology and heterogeneous engineering: the case of Portuguese expansion. Soc. Constr. Technol. Syst. New Dir. Sociol. Hist. Technol. **1**, 1–134 (1987)
41. Suchman, L.: Human-Machine Reconfigurations: Plans and Situated Actions. Cambridge University Press, Cambridge (2007)
42. Faraj, S., Azad, B.: The materiality of technology: an affordance perspective. In: Materiality and Organizing: Social Interaction in a Technological World, pp. 237–258 (2012)
43. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)
44. Majchrzak, A., Faraj, S., Kane, G.C., Azad, B.: The contradictory influence of social media affordances on online communal knowledge sharing. J. Comput.-Mediat. Commun. **19**, 38–55 (2013)
45. Markus, M.L., Silver, M.: A foundation for the study of IT effects: a new look at desanctis and poole's concepts of structural features and spirit. J. Assoc. Inf. Syst. **9**, 609–632 (2008)
46. Zammuto, R.F., Griffith, T.L., Majchrzak, A., Dougherty, D.J., Faraj, S.: Information technology and the changing fabric of organization. Organ. Sci. **18**, 749–762 (2007)
47. Treem, J.W., Leonardi, P.M.: Social media use in organizations: exploring the affordances of visibility, editability, persistence, and association. SSRN Electron. J. **36**, 143–189 (2012)

Records Systems and Information Systems: Connecting in Organizations

Sherry L. Xie[1,2,3(✉)] and Guanyan Fan[3]

[1] Key Laboratory of Data Engineering and Knowledge Engineering,
Ministry of Education of China, Beijing 100872, China
sherrylx@outlook.com

[2] Center for Electronic Records Management Research, Beijing 100872, China

[3] School of Information Resource Management, Renmin University of China,
59 Zhongguancun Ave, Haidian District, Beijing 100872, China

Abstract. The field of Information Systems (ISs) has long been recognized, so has Enterprise Information Systems (EISs), a field close to it. Long existing also in organizations or enterprises is the field of records management, now predominantly digital records management (DRM), which shares the many goals of ISs and EISs in supporting the operation and success of organizations. While the DRM field recognized rather early in its battle to digital records challenges that the need to establish formal collaborations with the ISs and EISs professions, it is still rare to spot today discussions regarding such collaborations in the general information and communication technology (ICT) literature. To help bridge the gap, this chapter introduces to the ISs and EISs professions one of the major developments of the international DRM field, that is, the records systems elaborated by the InterPARES (International Research on Permanent and Authentic Records in Electronic Systems) project, for the purpose of invoking further discussions.

Keywords: Records systems · Information systems
Enterprise information systems · InterPARES · Chain of preservation model

1 Introduction

Records systems and information systems (ISs) including Enterprise Information Systems (EISs) typically co-exist in organizations, in particular in those who operate under rigorous regulatory frameworks. While lacking universally agreed definitions, ISs and EISs can generally be understood as configurations of information and communication technologies (ICTs) that are deployed in organizations for the purpose of supporting organizations' accomplishment of business goals [1]. Sharing the same purpose, the organizational program for records management – now predominantly digital records management (DRM) – claims its establishment in organizations by facilitating the achievement of operational efficiency and effectiveness as well as legal compliance. The DRM field recognized rather early in its battle to digital records challenges the need to establish formal collaborations with the ISs and EISs professions for devising ISs functional requirements and for developing long term preservation

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, pp. 363–394, 2019.
https://doi.org/10.1007/978-3-319-99701-8_18

strategies, it is still rare, however, to spot discussions regarding such collaborations in the information and communication technology (ICT) literature today. Working with the understanding that both ISs/EISs and DRM are charged with the responsibilities of serving the business needs of their sponsoring organizations, this chapter, a substantive extension of the conference paper Organizational Records Systems - An Alternative View to (Enterprise) Information Systems [2], analyzes the relationships between records systems and information systems utilizing, as a representative case, the development of the InterPARES (International Research on Permanent and Authentic Records in Electronic Systems) project. To our knowledge, despite that research in both the fields of information systems and records systems abound, the analysis as conducted in this chapter appears to be the first of its kind. This chapter consists of 5 sections: Introduction (this section), Record(s) and Information, Information Systems and Records Systems, InterPARES Records Systems, and Conclusions. Due to the highly conceptual nature of the COP model, efforts are made to streamline the presenting process, incorporating approaches of general account, graph depiction, and definitions as explanations.

2 Record(s) and Information

The relationship between record(s) and information was once clear. In the paper world, or, to be more precise, in the world where information systems were not the primary platform for organizations to conduct their businesses, the use of the term information suggested informal and the use of the term record(s) suggested formal. An organizational record consisted of two major parts, content and documentary form [3] and it was recognized by the legal and judicial systems in which its creating organization was operating. By the long established theoretical, methodological, and analytical frameworks, the management of record(s) ensured record(s) reliability, authenticity, and trustworthiness. Collectively, organizational records enabled the establishment and continued existence of organizations, supported their functions and activities, and provided them with foundations on which progression and protection could be built. In this context, (written) information constituted the part of content of a record, and its organization, or the shape of it, conformed to the documentary form of the record. The existence and importance of information, therefore, was manifested in and through records.

Digital (computer) technologies blurred the clarity of this relationship in a truly disruptive way. The term information, along with the term data, was given a new, much more active life by business oriented information systems, which first entered into organizations in the form of "electronic data processing" (EDP) [4, 5] and then, of business applications backed up with (relational) database technologies [5, 6]. With the nonstop advancing of digital technologies, information systems become increasingly wide-spread, common in organizations, and information is now "assets" and "lifeblood" for organizations. The once informal term started to obtain a formal status, in either organizational policies and/or governmental regulations. For example, the Australian Federal Government [7, 8] acknowledges formally that information is "knowledge communicated and received" and the U.S. Federal Government [9] defines information in its government-wide policy as "any communication or representation of knowledge such as facts, data, or opinions in any medium or form, including textual, numerical, graphic, cartographic, narrative, electronic, or audiovisual forms".

Existing in the same digital world, record(s), however, are persistently viewed by non-records professions as in unbreakable bonds with analog formats and irrelevant to information systems – despite the fact that the records field has been transforming the management of analogue records into that for electronic/digital records since the early 1970s when the pioneering U.S. National Archives and Records Administration (NARA) started to handle data files in the form of punch cards. To the deployment of new information technologies in organizations, the management of records is always an afterthought. For example, the system of managing electronic document(s), which were loosely used to refer to both electronically or digitally captured paper documents, entered organizations without making any reference to organizational records management, and email, a typical representative of the digital disruptive power, remain till today to be a long reach of the program of organizational records management. This reality raises questions such as whether information can replace records in organizations, whether the notion of information as records content fitting into the documentary form of records can hold still true, and what relationships records now have with information in information systems, etc. It is unlikely that information is going to replace records in organizations, at least not in the current time. Organizational records in the digital world imply still authority and maintain still their legal and judicial status. For this reason alone, the blurred relationships between records and information could result in ineffective operation or regulatory violation. They, therefore, warrant to be cleared up.

To clear up the relationships between information and records in digitalized organizations proved to be challenging. Organizations either do not have formal definitions for information or avoid to define it altogether. For example, the Canadian Federal Government, while has in place a government-wide policy on information management, provides no definitions for information – the subject of the Policy on Information Management. Moreover, this policy treats records management as a constituent part of information management, yet its subordinate policy instrument, the Directive on Recordkeeping, introduces the concepts of information resources and information resources of business value, and treats them as the subject of the Directive – despite the title of the Directive speaks to "Recordkeeping". Although the three concepts of record, information resources, and information resources of business value are formally defined (record is defined in both the Policy and the Directive with identical words), the conceptual relationships between them, and with information, remain problematic, if not entirely impossible, to be clearly identified. The three definitions are listed below for illustration purpose:

- record: for the purpose of this policy, records are information created, received, and maintained by an organization or person for business purposes, legal obligations, or both, regardless of medium or form [10];
- information resources: any documentary material produced in published and unpublished form regardless of communications source, information format, production mode or recording medium. Information resources include textual records

(memos, reports, invoices, contracts, etc.), electronic records (e-mails, databases, internet, intranet, data etc.), new communication media (instant messages, wikis, blogs, podcasts, etc.), publications (reports, books, magazines), films, sound recordings, photographs, documentary art, graphics, maps, and artefacts [11];

- information resources of business value: published and unpublished materials, regardless of medium or form, that are created or acquired because they enable and document decision-making in support of programs, services and ongoing operations, and support departmental reporting, performance and accountability requirements (Ibid);

Some organizations do have a formal definition for information, such as the U.S. Federal Government introduced above. The issue with this case is that once the general definition of information is applied to the specific organizational setting, the distinguishing ability of the definition starts to fade. For example, apart from the definition of information, the OMB Circular No. A-130 Management of Federal Information Resources defines also federal information and public information, with the former being "information created, collected, processed, maintained, disseminated, disclosed, or disposed of by or for the Federal Government, in any medium or form" and the latter being "any information, regardless of form or format, that an agency discloses, disseminates, or makes available to the public". Because both definitions are in the context of agency operation and legal compliance, it becomes challenging to distinguish information from records in this setting. According to Title 44 of the U.S.C., § 3301, records "includes all recorded information, regardless of form or characteristics, made or received by a Federal agency under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the United States Government or because of the informational value of data in them" [12]. Further, for the purpose of emphasizing the relevance of record(s) to the digital world, the second part of the definition explains that "the term 'recorded' information includes all traditional forms of records, regardless of physical form or characteristics, including information created, manipulated, communicated, or stored in digital or electronic form". Federal information and federal record(s), therefore, appear to largely overlap with each other in this context.

3 Information Systems and Records Systems

The ISs field came to be recognized in the 1960s [13] and has ever since been advancing. Although missing in definitions of ISs universally accepted wordings, the linkage between ISs and organizations appears to always exists, either explicitly or implicitly. Internally, ISs are implemented to support the alternatively called back-office functions such as finance management, property management, and human capital management, and for functions facing customers or suppliers such as customer relationship management (CRM), supply chain management (SCM), sales, and marketing etc., ISs are designed with specifications tailored to address their functional requirements. The linkage between ISs and organizations became completely explicit when the notion of Enterprise Information Systems (EISs) started to emerge in the 1980s [14]. While choosing to use the term enterprise, EISs has never limited its applications to only commercial organizations. Like ISs, the phrase enterprise information systems (EISs) does not appear to be defined with consensus, and interpretations in the ICT field vary. For example, in the editorial of the inaugural issue of the Enterprise Information Systems journal, EISs was introduced as the equivalent to Enterprise Resource Planning (ERP) [14], yet in Enterprise Information Systems and Implementing IT Infrastructures: Challenges and Issues, EISs "comprises of information systems such as enterprise resource planning (ERP), supply chain management (SCM), customer relationship management (CRM) and e-commerce" [15]. Nonetheless, EISs is generally seen as, with reference to the ineffectiveness and failures of the earlier ISs implementations, the more logical and intelligent response to organizations' ISs needs. Advanced ISs and EISs are expected to better address the increasingly complex global environment and the increasingly integrative nature of business operations.

The most basic promises of EISs lies with integration, be it business process integration, system integration, data/information integration, or all the above. With integration, organizations can function as a whole: business processes can be streamlined, information silos can be bridged, and data integrity can be better ensured. As a result, wastes of ICT investment can be largely reduced (if not entirely avoided), employee resistance to new technologies can be minimized, and information can be available in real or near real time and be shared as needed irrespective of boundaries of business units and/or organizations. Ultimately, enterprises can be leaner, more agile, efficient and effective. Together, ISs and EISs promise to offer many qualities desired for organizations to achieve business sustainability and competitive advantages.

Although conceptually overlapping in the digital world in terms of their subjects, the DRM field remains to be independent from the field of information management in general, at least in the setting of government organizations. According to Title 44 U.S.C. § 2901(2), records management means "the planning, controlling, directing, organizing, training, promoting, and other managerial activities involved with respect to records creation, records maintenance and use, and records disposition in order to achieve adequate and proper documentation of the policies and transactions of the Federal Government and effective and economical management of agency operations" [16]. By the definition of records introduced in the previous session, records management logically include digital records management. To achieve the goals of DRM, both policies and resources are needed, and digital technology is now one of such indispensable resources. The application of digital technologies to records management resulted in, most typically, the system for managing digital records, or records system. In the field of records management, records system can be understood either narrowly or broadly. The narrow or specific view of records system focuses on the digital systems that manage unstructured records, such as those produced by Microsoft Word Suite and the all kinds of email applications. The broad view, on the other hand, addresses the entire spectrum of records management activities, regardless, thus, where the records exist and how they look like. A records system in the narrow view is indeed one type of business information systems because, as suggested by the above definition, records management or DRM is a business function in organizations and a technological system handling its activities is as the same as those information systems that handle any other business functions. For the purpose of this chapter, the broad view of records system is adopted and the InterPARES' development on records systems is chosen for the purpose of illustration and discussion.

4 InterPARES Records Systems

4.1 The InterPARES Project

The InterPARES project has been running for consecutively 16 years, in the form of four distinguishable yet interrelated phases, that is, the completed InterPARES I–III and the currently running InterPARES Trust (2013–2018). InterPARES phase I was built on the influential findings of the UBC-MAS Project, entitled Preservation of the Integrity of Electronic records, which had run from 1995 to 1997. The UBC-MAS Project, as one of the true pioneering digital records projects international, not only produced renewed conceptual knowledge pertinent to records and documents in digital formats, but also the first blueprint internationally for digital technologies to be meaningfully applied to digital records management. The DoD5015.02-STD, entitled "Electronic Records Management Software Applications Design Criteria Standard" (most recent edition issued on April 25, 2007), was a joint production of the UBC-MAS project and the United States Department of Defense Records Management Task Force, which specifies the configuration of ICTs for the management of digital records. By nature, such records management applications are ISs, and they provide ISs functionalities to facilitate organizations' conduct of activities regarding digital records management.

Initiated in 2001 and completed in 2012, InterPARES I–III extended its inquiries into digital records in more than one ways. With hundreds of researchers and graduate research assistants, the project had investigated a variety of research topics in a broad realm of domains, including, for example, digital arts, electronic government, and electronic science, against the technological backdrop of databases, document management system, and dynamic, interactive and experiential systems that heavily rely on network technologies. As a result, the project had developed an organization/enterprise-wide understanding of ISs, EISs, and captured it in one of its major products, that is, the Chain of Preservation (COP) model.

4.2 The COP Model

The name of the COP model points to the ultimate objective of the InterPARES project, that is, to ensure long-term or permanent preservation and accessibility of digital records – a representative reflection of the mission of the records community [17]. For the records community, this model serves both the professionals of the fields of DRM and digital archival administration in that both work with the same materials, that is, records, and the latter requires the former as managerial foundation. Together, these two professions complete the lifecycle management of records, with DRM disposing of, at any given time, records that are no longer needed and the archival administration entities providing custody and assess to significant records that require long-term or permanent preservation, upon transferring from DRM. Specifically, the COP model presents the lifecycle management in three types of systems: record-making system, record-keeping system, and record-preservation system.

Record-Making System. To apply the broad view of records system, record-making system in the context of the InterPARES project means “a set of rules governing the making of records, and the tools and mechanisms used to implement these rules”. Correspondingly, the activity of record-making encompasses “the whole of the principles, policies, rules and strategies that controls the process of creating records from made or received documents”. It needs to point out that the concept of document here is discipline relevant and therefore, should not be interpreted as a synonym to records as commonly found in everyday communications. As established by the findings of the second phase of InterPARES, the concept of document now covers those in digital formats, and can be broadly categorized as static, interactive, and dynamic/interactive ones. These digital documents have a convertible relationship with digital records, that is, when certain conditions are met, digital documents can become digital records although the formats may not be exactly the same [18, 19].

With its goal being the provision of overall control and co-ordination of the activities that it contains, the Record-Making System is designed to consist of 3 major activities Monitor Performance of Record-Making System (A2.1), Manage Making & Receipt of Records (A2.2), and Manage Setting Aside of Completed Records (A2.3), as depicted by Fig. 1 [17]. In Fig. 1 (also as in all figures that follow), arrows from the left side represent input information and arrows from the right side represent output information, which may also be input information for the next activity depending on the specific situation. The arrows from top represent constrains to the activity and the arrows from bottom represent facilities to the activity. The tunnel around (some of) the arrows indicates that the information is relevant not only to the present activity but also to all of its sub-activities. The outputs, therefore, are the results of a comprehensive integration and synthesizing of all of these arrows.

A2.1 is an activity without sub-activities (indicated in the figures by the short oblique line at the upper-left corner) and its job is to assess the efficacy of the performance of the record-making system. Specifically, it is required to analyze performance reports on the operation of each of the record-making system’s sub-activities and issue activity directives and information on the performance of the system. The records it generates is kept for the use in continued maintenance of the entire COP. Both A2.2 and A2.3 are activities with sub-activities, with A2.2 aiming to provide overall control and co-ordination of document and record making and receipt activities and A2.3 aiming to provide overall control and co-ordination of the transfer of executed or completed records to the recordkeeping system.

Figure 2 [17] depicts A2.2, showing its 5 sub-activities of Make Documents, Capture Documents, Identify Documents, Declare Records, and Execute Records. It needs to point out that the term creator in this context refers to the records-creating organization as opposed to individual authors to a document or record. Table 1 provides definitions for these sub-activities.

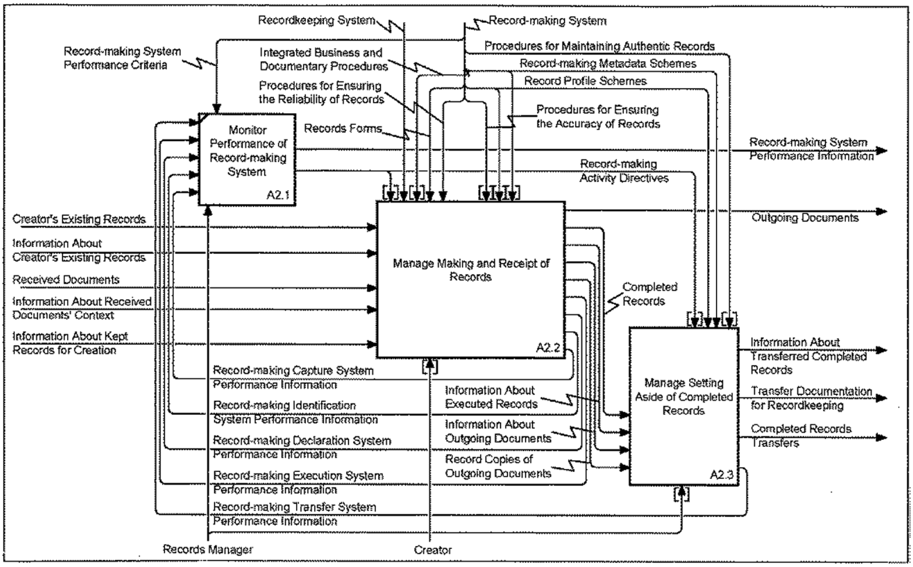

Recordkeeping System
Record-making System
Record-making System Performance Criteria
Integrated Business and Documentary Procedures
Procedures for Maintaining Authentic Records
Record-making Metadata Schemes
Record Profile Schemes
Procedures for Ensuring the Reliability of Records
Procedures for Ensuring the Accuracy of Records
Records Forms
Monitor Performance of Record-making System
A2.1
Record-making System Performance Information
Record-making Activity Directives
Creator's Existing Records
Information About Creator's Existing Records
Received Documents
Information About Received Documents' Context
Information About Kept Records for Creation
Manage Making and Receipt of Records
A2.2
Outgoing Documents
Completed Records
Manage Setting Aside of Completed Records
A2.3
Information About Transferred Completed Records
Transfer Documentation for Recordkeeping
Completed Records Transfers
Record-making Capture System Performance Information
Record-making Identification System Performance Information
Record-making Declaration System Performance Information
Record-making Execution System Performance Information
Record-making Transfer System Performance Information
Information About Executed Records
Information About Outgoing Documents
Record Copies of Outgoing Documents
Records Manager
Creator

Fig. 1. Record-making system.

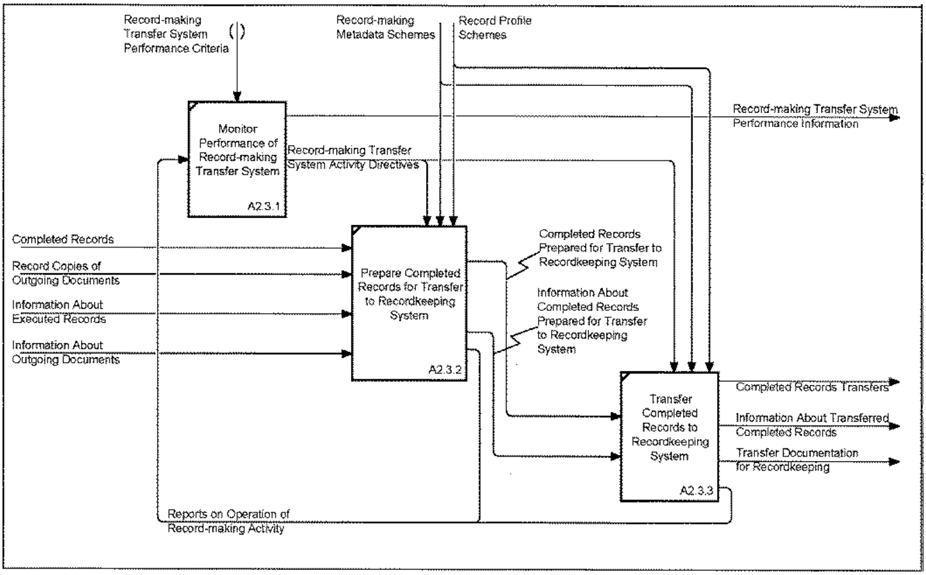

Record-making Transfer System Performance Criteria
Record-making Metadata Schemes
Record Profile Schemes
Monitor Performance of Record-making Transfer System
A2.3.1
Record-making Transfer System Performance Information
Record-making Transfer System Activity Directives
Completed Records
Record Copies of Outgoing Documents
Information About Executed Records
Information About Outgoing Documents
Prepare Completed Records for Transfer to Recordkeeping System
A2.3.2
Completed Records Prepared for Transfer to Recordkeeping System
Information About Completed Records Prepared for Transfer to Recordkeeping System
Transfer Completed Records to Recordkeeping System
A2.3.3
Completed Records Transfers
Information About Transferred Completed Records
Transfer Documentation for Recordkeeping
Reports on Operation of Record-making Activity

Fig. 2. Making and receiving of records.

Figure 3 [17] depicts A2.3, showing its 3 sub-activities of Monitor Performance of Record-making Transfer System, Prepare Completed Records for Transfer to Recordkeeping System, and Transfer Completed Records to Recordkeeping System. Table 2 provides definitions for these sub-activities.

Table 1. Definitions for A2.2 sub-activities.

Activity	Definition
A2.2.1 Make documents	To compile digital information in a syntactic manner in accordance with the specifications of the creator's documentary forms, integrated business and documentary procedures and record-making access privileges
A2.2.2 Capture documents	To record and save (i.e., affix to a digital medium in a stable syntactic manner) particular instantiations of incoming external documents or internal documents made by the creator in the record-making system in accordance with the specifications of the creator's integrated business and documentary procedures and record-making access privileges
A2.2.3 Identify documents	To attach to each document identity metadata that convey the action in which the document participates and its immediate context
A2.2.4 Declare records	To intellectually set aside records by assigning classification codes from the classification scheme to made or received documents and adding these codes to the identifying metadata and by assigning to the documents registration numbers based on the registration scheme, and adding these numbers to the identifying metadata
A2.2.5 Execute records	To attach to each record metadata that convey information related to, and actions taken during the course of, the formal execution phase of the administrative procedure in which the record participates, which may also involve transmitting documents to external physical or juridical persons and making record copies of the sent documents

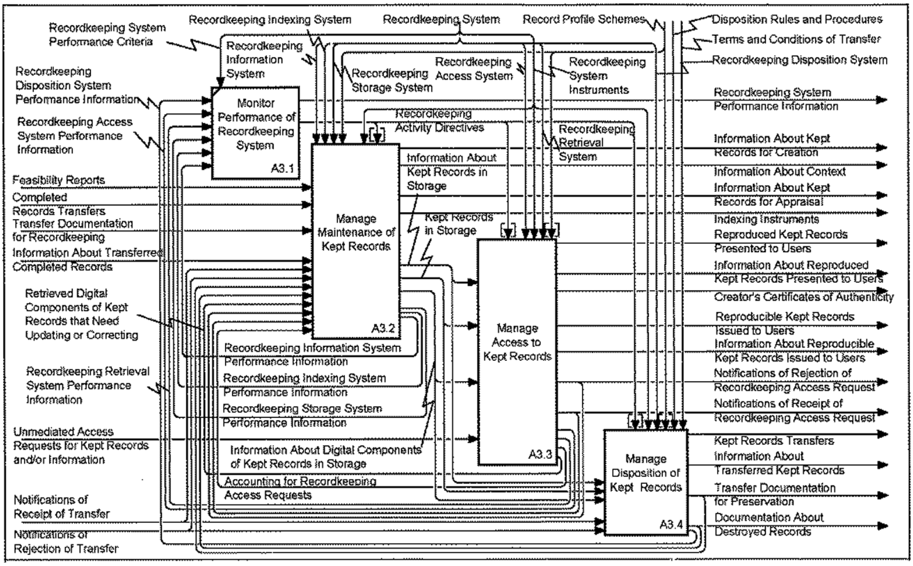

Recordkeeping System Performance Criteria
Recordkeeping Indexing System
Recordkeeping System
Record Profile Schemes
Disposition Rules and Procedures
Terms and Conditions of Transfer
Recordkeeping Disposition System
Recordkeeping Information System
Recordkeeping Storage System
Recordkeeping Access System
Recordkeeping System Instruments
Recordkeeping Disposition System Performance Information
Recordkeeping Access System Performance Information
Monitor Performance of Recordkeeping System
A3.1
Recordkeeping Activity Directives
Recordkeeping Retrieval System
Feasibility Reports
Completed Records Transfers
Transfer Documentation for Recordkeeping
Information About Transferred Completed Records
Manage Maintenance of Kept Records
A3.2
Information About Kept Records in Storage
Kept Records in Storage
Retrieved Digital Components of Kept Records that Need Updating or Correcting
Recordkeeping Information System Performance Information
Recordkeeping Indexing System Performance Information
Recordkeeping Retrieval System Performance Information
Recordkeeping Storage System Performance Information
Unmediated Access Requests for Kept Records and/or Information
Information About Digital Components of Kept Records in Storage
Accounting for Recordkeeping
Access Requests
Notifications of Receipt of Transfer
Notifications of Rejection of Transfer
Manage Access to Kept Records
A3.3
Manage Disposition of Kept Records
A3.4
Recordkeeping System Performance Information
Information About Kept Records for Creation
Information About Context
Information About Kept Records for Appraisal
Indexing Instruments
Reproduced Kept Records Presented to Users
Information About Reproduced Kept Records Presented to Users
Creator's Certificates of Authenticity
Reproducible Kept Records Issued to Users
Information About Reproducible Kept Records Issued to Users
Notifications of Rejection of Recordkeeping Access Request
Notifications of Receipt of Recordkeeping Access Request
Kept Records Transfers
Information About Transferred Kept Records
Transfer Documentation for Preservation
Documentation About Destroyed Records

Fig. 3. Setting aside of completed records.

Table 2. Definitions for A2.3 sub-activities.

Activity	Definition
A2.3.1 Monitor performance of record-making transfer system	To assess the efficacy of the performance of the record-making transfer system by analyzing reports on the operation of record-making activities, and issue activity directives for transfer activities and issue information on the performance of the record-making transfer system for use in continued maintenance of the record-making system
A2.3.2 Prepare completed records for transfer to recordkeeping system	To attach to completed records integrity and related metadata that convey information related to, and actions taken during the course of, managing the records for records management purposes prior to setting them aside in the recordkeeping system; compile information about the records that is needed to meet all transfer information requirements; and ensure that the records are in the proper format for transfer to the recordkeeping system as prescribed by recordkeeping system rules and procedures and technological requirements
A2.3.3 Transfer completed records to recordkeeping system	To send or transmit completed records prepared for transfer to the office responsible for the recordkeeping function with the accompanying documentation necessary for recordkeeping

Record-Keeping System. Like record-making system, the InterPARES record-keeping system goes beyond the narrow view of technological system and refers to the whole set of "rules governing the storage, use, maintenance and disposition of records and/or information about records, and the tools and mechanisms used to implement these rules". Figure 4 [17] depicts such a system, showing the activities designed to achieve the goal of providing overall control and co-ordination of activities in the recordkeeping system, including records storage, retrieval and access, disposition, and monitoring of the performance of the record-keeping system.

In this context, the operation of record-keeping requires a creator to identify the principles, policies, rules and strategies necessary to establish and maintain administrative, intellectual, and physical control over its records, and the activities of A3 Manage Records in a Recordkeeping System are designed to facilitate such an operation. A3 consists of 4 activities: A3.1 Monitor Performance of Recordkeeping System, A3.2 Manage Maintenance of Kept Records, A3.3 Manage Access to Kept Records, and A3.4 Manage Disposition of Kept Records, with the later 3 containing sub-activities. Table 3 provides definitions of these activities, where preserver refers to the entity that is given formally the authority and responsibility of managing records in a permanent manner.

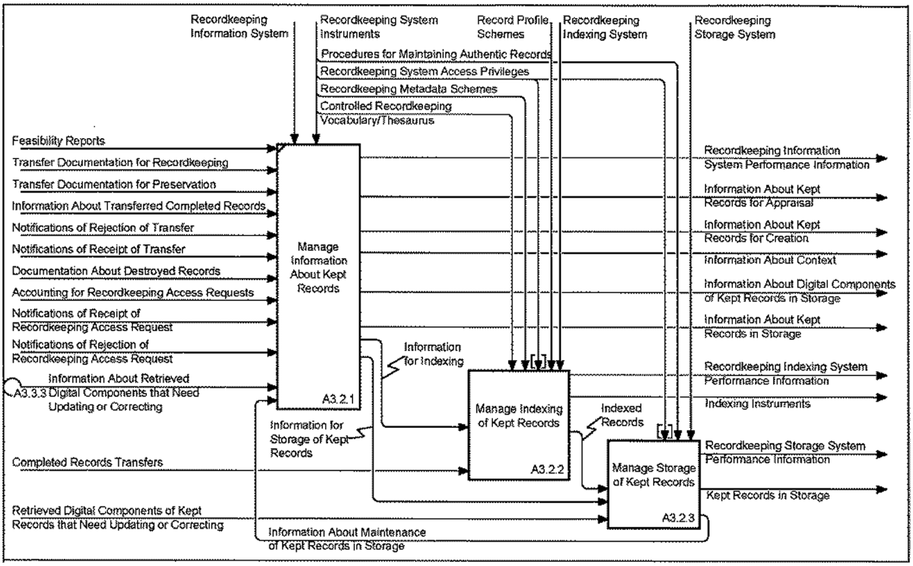

Recordkeeping Information System
Recordkeeping System Instruments
Record Profile Schemes
Recordkeeping Indexing System
Recordkeeping Storage System
Procedures for Maintaining Authentic Records
Recordkeeping System Access Privileges
Recordkeeping Metadata Schemes
Controlled Recordkeeping Vocabulary/Thesaurus
Feasibility Reports
Transfer Documentation for Recordkeeping
Transfer Documentation for Preservation
Information About Transferred Completed Records
Notifications of Rejection of Transfer
Notifications of Receipt of Transfer
Documentation About Destroyed Records
Accounting for Recordkeeping Access Requests
Notifications of Receipt of Recordkeeping Access Request
Notifications of Rejection of Recordkeeping Access Request
Information About Retrieved
A3.3.3 Digital Components that Need Updating or Correcting
Manage Information About Kept Records
A3.2.1
Information for Indexing
Information for Storage of Kept Records
Completed Records Transfers
Retrieved Digital Components of Kept Records that Need Updating or Correcting
Information About Maintenance of Kept Records in Storage
Manage Indexing of Kept Records
A3.2.2
Indexed Records
Manage Storage of Kept Records
A3.2.3
Recordkeeping Information System Performance Information
Information About Kept Records for Appraisal
Information About Kept Records for Creation
Information About Context
Information About Digital Components of Kept Records in Storage
Information About Kept Records in Storage
Recordkeeping Indexing System Performance Information
Indexing Instruments
Recordkeeping Storage System Performance Information
Kept Records in Storage

Fig. 4. Record-keeping system.

Table 3. Definitions for A3 sub-activities.

Activity	Definition
A3.1 Monitor performance of recordkeeping system	To assess the efficacy of the performance of the recordkeeping system by analyzing performance reports on the operation of recordkeeping sub-system activities, and issue activity directives for recordkeeping activities and information on the performance of the recordkeeping system for use in continued maintenance of the chain of preservation framework
A3.2 Manage maintenance of kept records	To provide overall control and co-ordination of the recordkeeping storage system and the records stored in the system by managing information about kept records and their digital components, placing the records in storage, maintaining the digital components and monitoring the performance of the storage system
A3.3 Manage access to kept records	To facilitate discovery of, and manage requests for, kept records and/or information about kept records, and monitor the performance of the recordkeeping access system
A3.4 Manage disposition of kept records	To provide overall control and co-ordination of records disposition activities, including monitoring the performance of the disposition system, processing disposition information and, in accordance with disposition activity directives and disposition rules and procedures, destroying kept records and/or preparing and transferring kept records to the designated preserver

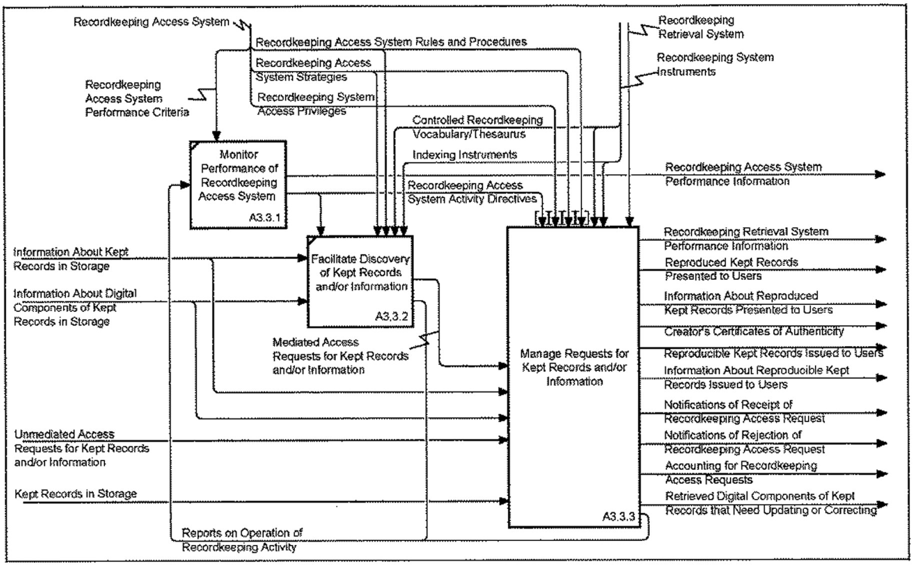

Recordkeeping Access System
Recordkeeping Access System Rules and Procedures
Recordkeeping Access System Strategies
Recordkeeping System Access Privileges
Recordkeeping Access System Performance Criteria
Recordkeeping Retrieval System
Recordkeeping System Instruments
Controlled Recordkeeping Vocabulary/Thesaurus
Indexing Instruments
Recordkeeping Access System Activity Directives
Monitor Performance of Recordkeeping Access System
A3.3.1
Facilitate Discovery of Kept Records and/or Information
A3.3.2
Manage Requests for Kept Records and/or Information
A3.3.3
Information About Kept Records in Storage
Information About Digital Components of Kept Records in Storage
Mediated Access Requests for Kept Records and/or Information
Unmediated Access Requests for Kept Records and/or Information
Kept Records in Storage
Reports on Operation of Recordkeeping Activity
Recordkeeping Access System Performance Information
Recordkeeping Retrieval System Performance Information
Reproduced Kept Records Presented to Users
Information About Reproduced Kept Records Presented to Users
Creator's Certificates of Authenticity
Reproducible Kept Records Issued to Users
Information About Reproducible Kept Records Issued to Users
Notifications of Receipt of Recordkeeping Access Request
Notifications of Rejection of Recordkeeping Access Request
Accounting for Recordkeeping Access Requests
Retrieved Digital Components of Kept Records that Need Updating or Correcting

Fig. 5. Maintaining kept records.

Figure 5 [17] depicts A3.2, showing its 3 sub-activities of A3.2.1 Manage Information About Kept Records, A3.2.2 Manage Indexing of Kept Records, and A3.2.3 Manage Storage of Kept Records. Table 4 provides definitions for these sub-activities.

Table 4. Definitions for A3.2 sub-activities.

Activity	Definition
A3.2.1 Manage information about kept records	To compile information about records in the recordkeeping system and about records maintenance activities and to provide overall control and co-ordination of that information for use in records appraisal activities by the preserver and in records indexing, storage, access and disposition activities by the creator
A3.2.2 Manage indexing of kept records	To provide overall control and co-ordination of records indexing activities, including monitoring the indexing system, indexing kept records and developing indexing instruments to help facilitate records discovery and retrieval
A3.2.2.1 Monitor performance of recordkeeping indexing system	To assess the efficacy of the performance of the recordkeeping indexing system by analyzing reports on the operation of recordkeeping activities, and issue activity directives for indexing activities and information on the performance of the indexing system for use in continued maintenance of the recordkeeping system

(*continued*)

Table 4. (*continued*)

Activity	Definition
A3.2.2.2 Index kept records	To establish and record access points for kept records within the context of a controlled recordkeeping vocabulary applied according to recordkeeping indexing system rules, procedures and strategies
A3.2.2.3 Develop indexing instruments	To prepare tools that facilitate discovery and retrieval of the records in the recordkeeping system, such as guides, inventories and indexes
A3.2.3 Manage storage of kept records	To provide overall control and co-ordination of the recordkeeping storage system and the records stored in the system by placing the records in storage, maintaining their digital components and monitoring the performance of the storage system
A3.2.3.1 Monitor performance of recordkeeping storage	To assess the efficacy of the performance of the recordkeeping storage system by analyzing reports on the operation of recordkeeping activities, and issue activity directives for storage activities and information on the performance of the recordkeeping storage system for use in continued maintenance of the recordkeeping system
A3.2.3.2 Place kept records in storage	To place the digital components of kept records and their metadata into storage in accordance with the procedures for maintaining authentic records and the actions prescribed by the recordkeeping storage system strategies, rules and procedures and activity directives
A3.2.3.3 Maintain records in recordkeeping storage system	To monitor the storage of kept records and their digital components and metadata, periodically back-up the recordkeeping storage system and, as necessary, correct problems with and update the digital components, and/or refresh storage media to ensure the records in the system remain accessible, legible and intelligible over time
A3.2.3.3.1 Monitor kept records in storage	To keep track of the condition and maintenance requirements of kept records and their digital components–more specifically, their digital components and metadata–and the media on which they are stored in the recordkeeping storage system to identify storage that needs backing-up, digital components and/or metadata that need correcting or updating and media that need refreshing; and to issue reports on maintenance activities
A3.2.3.3.2 Back-up recordkeeping storage system	To routinely make a copy of all digital content in the recordkeeping storage system, including the operating system, the software applications and all digital objects in the system, for the purpose of recovery in the event of a disaster resulting in system failure or corruption, and record information about these back-up activities

(*continued*)

Table 4. (*continued*)

Activity	Definition
A3.2.3.3.3 Correct problems with kept records in storage	To take the actions prescribed by the recordkeeping storage system strategies, rules and procedures and activity directives, in accordance with the procedures for maintaining authentic records, to eliminate problems in storage, and record information about these correction activities
A3.2.3.3.4 Update kept records in storage	To carry out conversion actions on the digital components of stored kept records in accordance with the procedures for maintaining authentic records and the actions prescribed by the recordkeeping storage system strategies, rules and procedures and activity directives, to ensure the records remain accessible, legible and intelligible over time (such as by migration, standardization or transformation to persistent form), and record information about these updating activities
A3.2.3.3.5 Refresh media for kept records in storage	To copy or transfer the digital components of kept records in storage from one medium to another, or otherwise ensure the storage medium remains sound, in accordance with the procedures for maintaining authentic records and the actions prescribed by the recordkeeping storage system strategies, rules and procedures and activity directives, and record information about these media refreshment activities

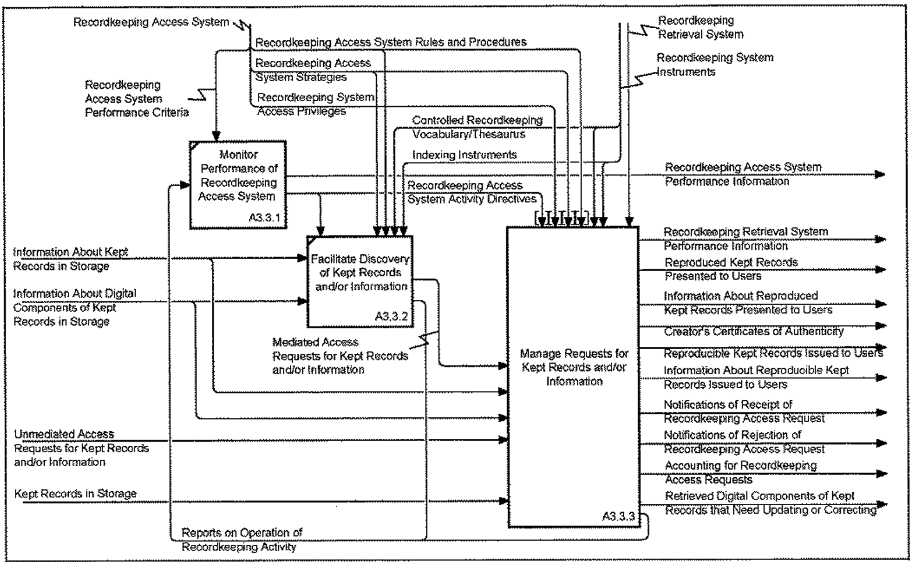

Recordkeeping Access System
Recordkeeping Access System Rules and Procedures
Recordkeeping Access System Strategies
Recordkeeping System Access Privileges
Recordkeeping Access System Performance Criteria
Recordkeeping Retrieval System
Recordkeeping System Instruments
Controlled Recordkeeping Vocabulary/Thesaurus
Indexing Instruments
Recordkeeping Access System Activity Directives
Monitor Performance of Recordkeeping Access System
A3.3.1
Recordkeeping Access System Performance Information
Information About Kept Records in Storage
Information About Digital Components of Kept Records in Storage
Facilitate Discovery of Kept Records and/or Information
A3.3.2
Mediated Access Requests for Kept Records and/or Information
Unmediated Access Requests for Kept Records and/or Information
Kept Records in Storage
Manage Requests for Kept Records and/or Information
A3.3.3
Reports on Operation of Recordkeeping Activity
Recordkeeping Retrieval System Performance Information
Reproduced Kept Records Presented to Users
Information About Reproduced Kept Records Presented to Users
Creator's Certificates of Authenticity
Reproducible Kept Records Issued to Users
Information About Reproducible Kept Records Issued to Users
Notifications of Receipt of Recordkeeping Access Request
Notifications of Rejection of Recordkeeping Access Request
Accounting for Recordkeeping Access Requests
Retrieved Digital Components of Kept Records that Need Updating or Correcting

Fig. 6. Access to kept records.

Table 5. Definitions for A3.3 sub-activities.

Activity	Definition
A3.3.1 Monitor performance of recordkeeping access system	To assess the efficacy of the performance of the recordkeeping access system by analyzing reports on the operation of recordkeeping activities, and issue activity directives for access activities and information on the performance of the recordkeeping access system for use in continued maintenance of the recordkeeping system
A3.3.2 Facilitate discovery of kept records and/or information	To provide authorized internal and external users access to, and assistance in the use of, the tools and resources necessary to support querying and searching for, and discovery of, information, records and/or records aggregates in the recordkeeping system suited to a particular inquiry or purpose
A3.3.3 Manage requests for kept records and/or information	To provide overall control and co-ordination of internal and external requests for access to records and/or information about kept records by processing access requests, retrieving digital components for requested records and/or information, verifying retrieved components and information and providing access to retrieved records and/or information
A3.3.3.1 Process requests for kept records and/or information	To register access requests for kept records and/or information, translate them, define request specifications, generate retrieval requests and account for any problems with processing requests
A3.3.3.1.1 Register recordkeeping access requests	To record registration information about received requests for access to kept records and/or information about the records and issue notifications of receipt to the persons requesting the records
A3.3.3.1.2 Retrieve information to process recordkeeping access requests	To gather the information, from indexing instruments, record profiles and other recordkeeping tools, needed to process access requests for kept records and/or information about records
A3.3.3.1.3 Generate recordkeeping retrieval requests	To translate access requests for kept records and/or information into requests to the recordkeeping storage and information systems for retrieval of the exact digital components and/or information required to fulfil the access requests
A3.3.3.1.4 Generate recordkeeping requests specifications	To issue instructions to the recordkeeping retrieval and access systems on how to fulfil requests for kept records and/or information about the records based on analyses of the requests and processing information in relation to recordkeeping access system strategies, rules and procedures (including procedures for maintaining authentic records) and access privileges

(*continued*)

Table 5. (*continued*)

Activity	Definition
A3.3.3.2 Retrieve requested kept records and/or information	To output copies of digital components of records, information about digital components of records, rendering information about records and/or content information about records retrieved from storage in the recordkeeping system in response to retrieval requests for components and/or information
A3.3.3.3 Verify retrieved kept records and/or information	To determine whether all components and information necessary to satisfy requests for kept records and/or information about kept records have been received and can be processed for output and, in cases where digital components are encountered that need updating or correcting, redirect them (or information about the problems encountered) to the maintenance function of the recordkeeping storage system
A3.3.3.4 Provide access to retrieved kept records and/or information	To fulfil access requests by either reconstituting the retrieved digital components of kept records and/or information in authentic form and presenting the manifested records or information to users, or by packaging the retrieved digital components with information about how to reconstitute and present the records and/or information with the appropriate extrinsic form and issuing the packaged materials to users, and account for the success or failure of either activity
A3.3.3.4.1 Reconstitute kept records and/or information	To link or assemble all the verified digital components of requested kept records and/or information about kept records as necessary to reproduce and present the records and/or information in authentic form and, if necessary, redact records and/or information to meet privacy and/or copyright requirements
A3.3.3.4.2 Manifest kept records and/or	To present copies of the reconstituted requested kept records and/or requested information about the records with the appropriate extrinsic form and with information about their relationships to one another (archival bond) and, if requested, produce a Certificate of Authenticity for the records copies
A3.3.3.4.3 Package kept records and/or information for output	To combine the digital components of the requested kept records and/or requested information about kept records with information on how to reconstitute and manifest the records or information with the appropriate extrinsic form

Figure 6 [17] depicts A3.3, showing its 3 sub-activities of A3.3.1 Monitor Performance of Recordkeeping Access System, A3.3.2 Facilitate Discovery of Kept Records and/or Information, and A3.3.3 Manage Requests for Kept Records and/or Information. Table 5 provides definitions of these sub-activities including those to A3.3.3.

Figure 7 [17] depicts A3.4, showing its 5 sub-activities of A3.4.1 Monitor Performance of Disposition System, A3.4.2 Identify Kept Records for Disposition, A3.4.3 Destroy Kept Records, A3.4.4 Prepare Kept Records for Transfer to Designated Preserver, and A3.4.5 Transfer Kept Records to Designated Preserver. Table 6 provides definitions for these sub-activities, none of which has sub-activities.

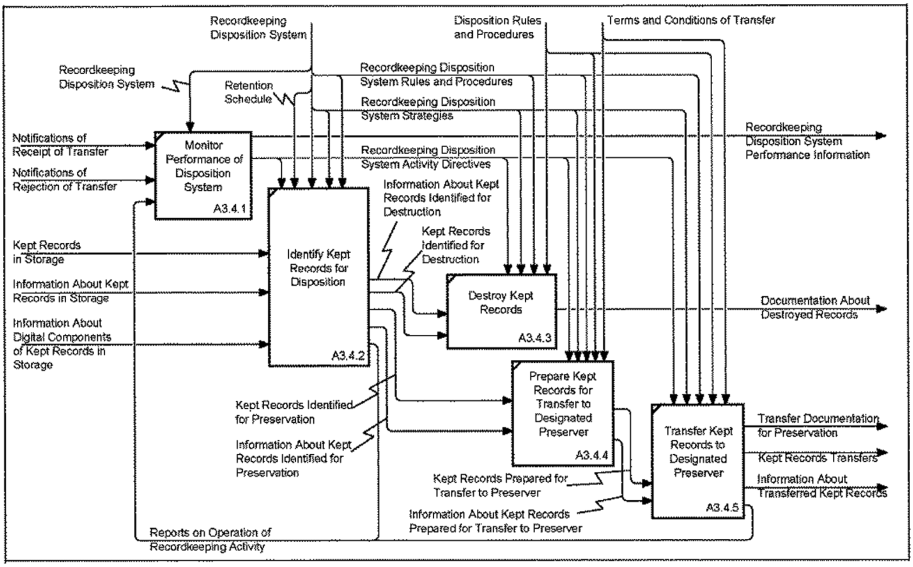

Recordkeeping Disposition System
Disposition Rules and Procedures
Terms and Conditions of Transfer
Recordkeeping Disposition System
Retention Schedule
Recordkeeping Disposition System Rules and Procedures
Recordkeeping Disposition System Strategies
Notifications of Receipt of Transfer
Notifications of Rejection of Transfer
Monitor Performance of Disposition System
A3.4.1
Recordkeeping Disposition System Activity Directives
Recordkeeping Disposition System Performance Information
Information About Kept Records Identified for Destruction
Kept Records Identified for Destruction
Kept Records in Storage
Information About Kept Records in Storage
Information About Digital Components of Kept Records in Storage
Identify Kept Records for Disposition
A3.4.2
Destroy Kept Records
A3.4.3
Documentation About Destroyed Records
Prepare Kept Records for Transfer to Designated Preserver
A3.4.4
Kept Records Identified for Preservation
Information About Kept Records Identified for Preservation
Transfer Kept Records to Designated Preserver
A3.4.5
Transfer Documentation for Preservation
Kept Records Transfers
Information About Transferred Kept Records
Kept Records Prepared for Transfer to Preserver
Information About Kept Records Prepared for Transfer to Preserver
Reports on Operation of Recordkeeping Activity

Fig. 7. Disposing of kept records

Record-Preserving System. Record-preserving system encompasses both rules governing the permanent intellectual and physical maintenance of acquired records and the tools and mechanisms needed to implement these rules. As the last type of records system in the COP model, it aims to provide overall control and co-ordination of activities in the permanent preservation system, including records appraisal and selection, acquisition, description, storage, retrieval and access, and monitoring of the performance of the permanent preservation system. To that end, the designated preserver is required to ical and technological stabilization and protecting the intellectual form of acquired/accessioned records, thus enabling records' "continuing, enduring, stable, lasting, uninterrupted and unbroken chain of preservation" [17]. Figure 8 [17] depicts the permanent preservation system, showing its 5 activities of A4.1 Monitor

Performance of Permanent Preservation System, A4.2 Appraise Records for Permanent Preservation, A4.3 Acquire Selected Records, A4.4 Preserve Accessioned Records, and A4.5 Output Records, with the later 4 containing their own sub-activities. Table 7 provides definitions for these activities.

Table 6. Definitions for A3.4 sub-activities.

Activity	Definition
A3.4.1 Monitor performance of disposition system	To assess the efficacy of the performance of the recordkeeping disposition system by analyzing reports on the operation of recordkeeping activities, and issue activity directives for disposition activities and information on the performance of the recordkeeping storage system for use in continued maintenance of the recordkeeping system
A3.4.2 Identify kept records for disposition	To identify records and information about records in the recordkeeping system earmarked either for destruction or transfer to the designated preserver, as determined by the creator's retention schedule
A3.4.3 Destroy kept records	To obliterate kept records, and information related to the records, identified for destruction and provide documentation about the records destroyed
A3.4.4 Prepare kept records for transfer to designated preserver	To attach to kept records integrity and related metadata about actions taken during the course of preparing the records for transfer to the designated preserver in accordance with the terms and conditions of transfer, and compile information about the records that is needed to meet all transfer information requirements
A3.4.5 Transfer kept records to designated preserver	To send or transmit kept records prepared for transfer to permanent preserver (or, as applicable, the office of the creator responsible for the permanent preservation function) with the accompanying documentation necessary for permanent preservation

Figure 9 [17] depicts the 4 sub-activities of A4.2: A4.2.1 Monitor Performance of Preservation Selection System, A4.2.2 Analyze Kept Records for Preservation, A4.2.3 Make Appraisal Decisions, and A4.2.4 Monitor Appraisal Decisions, with A4.2.2 showing to contain its own sub-activities. Table 8 provides definitions for these sub-activities including those to A4.2.2.2, A4.2.2.3, and A4.2.2.2.2.

Figure 10 [17] depicts the 3 sub-activities of A4.3: A4.3.1 Monitor Performance of Preservation Acquisition System, A4.3.2 Process Records Transfers, and A4.3.3 Accession Records, with A4.3.2 showing to contain its own sub-activities. Table 9 provides definitions for the sub-activities.

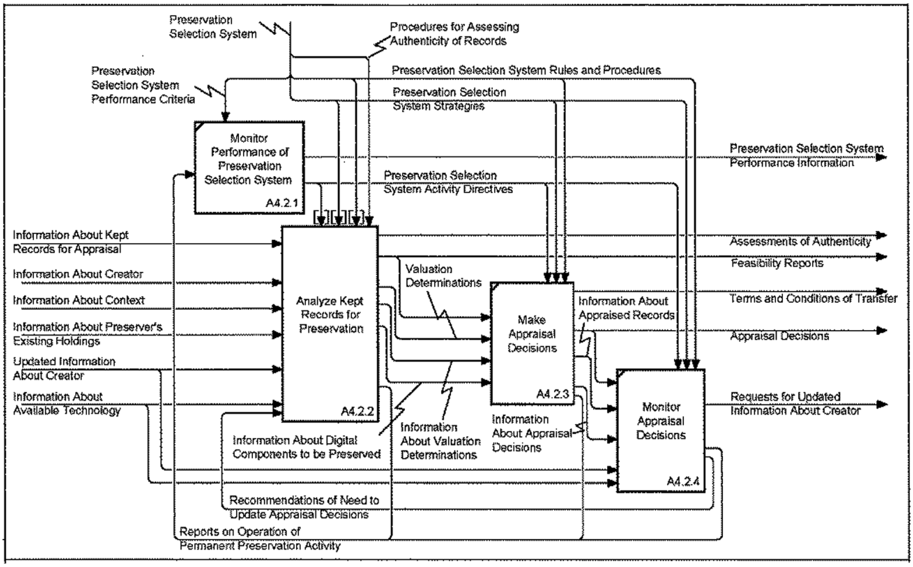

Preservation Selection System
Procedures for Assessing Authenticity of Records
Preservation Selection System Performance Criteria
Preservation Selection System Rules and Procedures
Preservation Selection System Strategies
Monitor Performance of Preservation Selection System
A4.2.1
Preservation Selection System Performance Information
Preservation Selection System Activity Directives
Information About Kept Records for Appraisal
Information About Creator
Information About Context
Information About Preserver's Existing Holdings
Updated Information About Creator
Information About Available Technology
Analyze Kept Records for Preservation
A4.2.2
Assessments of Authenticity
Feasibility Reports
Valuation Determinations
Terms and Conditions of Transfer
Make Appraisal Decisions
A4.2.3
Information About Appraised Records
Appraisal Decisions
Monitor Appraisal Decisions
A4.2.4
Requests for Updated Information About Creator
Information About Digital Components to be Preserved
Information About Valuation Determinations
Information About Appraisal Decisions
Recommendations of Need to Update Appraisal Decisions
Reports on Operation of Permanent Preservation Activity

Fig. 8. Record-preserving system/records permanent preservation system.

Table 7. Definitions for A4 sub-activities.

Activity	Definition
A4.1 Monitor performance of permanent preservation system	To assess the efficacy of the performance of the permanent preservation system by analyzing performance reports on the operation of permanent preservation sub-system activities, and issue activity directives for preservation activities and information on the performance of the permanent preservation system for use in continued maintenance of the chain of preservation framework
A4.2 Appraise records for permanent preservation	To make appraisal decisions by compiling information about kept records and their context, assessing their value, and determining the feasibility of their preservation; and to monitor appraised records and appraisal decisions to identify any necessary changes to appraisal decisions over time
A4.3 Acquire selected records	To bring records selected for permanent preservation into the custody of the preserver by registering and verifying transfers, confirming the feasibility of preservation, and accessioning the records or rejecting transfers if they are inadequate
A4.4 Preserve accessioned records	To manage information about, and the description and storage of, records acquired for permanent preservation
A4.5 Output records	To facilitate discovery of records and/or information about records in the permanent preservation system, manage requests for preserved records and/or information about the records and monitor the performance of the permanent preservation access system

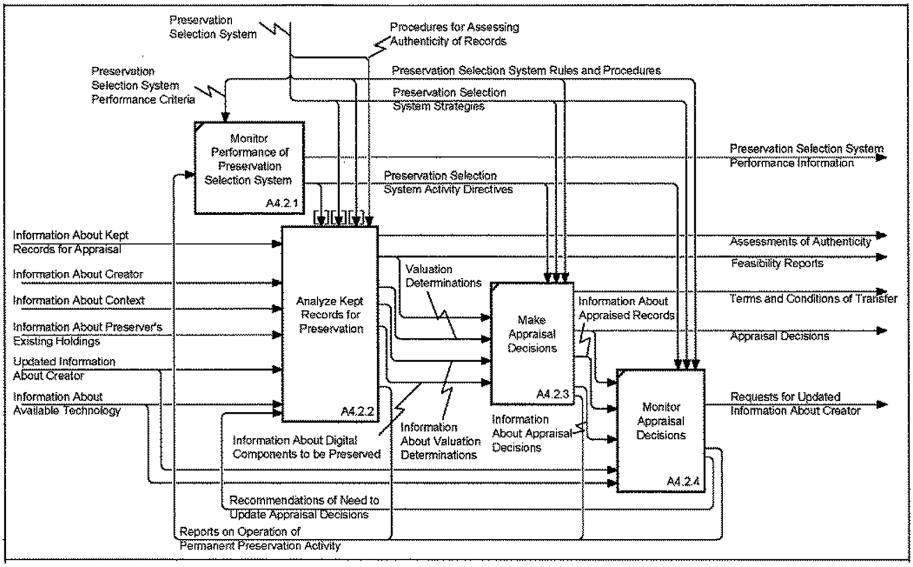

Preservation Selection System
Procedures for Assessing Authenticity of Records
Preservation Selection System Performance Criteria
Preservation Selection System Rules and Procedures
Preservation Selection System Strategies
Monitor Performance of Preservation Selection System
A4.2.1
Preservation Selection System Performance Information
Preservation Selection System Activity Directives
Information About Kept Records for Appraisal
Information About Creator
Information About Context
Information About Preserver's Existing Holdings
Updated Information About Creator
Information About Available Technology
Analyze Kept Records for Preservation
A4.2.2
Assessments of Authenticity
Feasibility Reports
Valuation Determinations
Make Appraisal Decisions
A4.2.3
Information About Appraised Records
Terms and Conditions of Transfer
Appraisal Decisions
Monitor Appraisal Decisions
A4.2.4
Requests for Updated Information About Creator
Information About Digital Components to be Preserved
Information About Valuation Determinations
Information About Appraisal Decisions
Recommendations of Need to Update Appraisal Decisions
Reports on Operation of Permanent Preservation Activity

Fig. 9. Appraising records for permanent preservation.

Table 8. Definitions for A4.2 sub-activities.

Activity	Definition
A4.2.1 Monitor performance of preservation selection system	To assess the efficacy of the performance of the permanent preservation selection system by analyzing reports on the operation of preservation activities, and issue activity directives for selection activities and information on the performance of the permanent preservation selection system for use in continued maintenance of the permanent preservation system
A4.2.2 Analyze kept records for preservation	To assess information concerning the kept records being appraised, including their contexts, value and preservation feasibility
A4.2.2.1 Analyze information about records	To collect, organize, record and assess relevant information from the kept records being appraised and about their juridical-administrative, provenancial, procedural, documentary and technological contexts
A4.2.2.2 Assess value of records	To analyze and judge: (1) the capacity of records being appraised to serve the continuing interests of their creator and society; and (2) the grounds for presuming the records to be authentic

(*continued*)

Table 8. (*continued*)

Activity	Definition
A4.2.2.2.1 Assess continuing value of records	To analyze and judge the capacity of records being appraised to serve the continuing interests of their creator and society
A4.2.2.2.2 Assess authenticity of records	To analyze and judge the grounds for presuming records being appraised to be authentic
A4.2.2.2.2.1 Compile evidence supporting the presumption of authenticity	To collect, organize and record evidence of the identity and integrity of records being appraised and about the procedural controls applied to them, to support the presumption of authenticity of those records
A4.2.2.2.2.2 Measure evidence against requirements for authentic records	To compare the evidence compiled about the identity, integrity and procedural controls of the records being appraised with the requirements for authentic records
A4.2.2.2.2.3 Verify authenticity	To use verification methods to determine the authenticity of records being appraised in cases where there is insufficient evidence to meet the requirements for presuming the authenticity of records
A4.2.2.2.3 Determine value of records	To establish the value of records being appraised based on assessments of their continuing value and their authenticity
A4.2.2.3 Determine feasibility of preservation	To identify the elements and digital components of the records being appraised, reconcile their preservation requirements with the preserver's current and anticipated preservation capabilities, and provide documentation about the digital components to be preserved and the feasibility of preservation
A4.2.2.3.1 Determine record elements to be preserved	To identify the necessary documentary components (e.g., record profile, attachments, annotations, etc.) and elements of form (e.g., author, date, subject line, etc.) of records to be preserved to determine which record elements must be preserved to protect the authenticity of those records
A4.2.2.3.2 Identify digital components to be preserved	To identify the digital components that manifest the record elements that need to be preserved to protect the authenticity of records selected for permanent preservation

(*continued*)

Table 8. (*continued*)

Activity	Definition
A4.2.2.3.3 Reconcile preservation requirements with preservation capabilities	To determine whether the digital components manifesting the record elements that need to be preserved to protect the authenticity of records selected for permanent preservation can in fact be preserved given the preserver's current and anticipated preservation capabilities
A4.2.3 Make appraisal decisions	To decide on and document the retention and disposition of records based on valuation and feasibility information, and to agree on and document the terms and conditions of transfer of the records to the preserver
A4.2.4 Monitor appraisal decisions	To keep track of appraisal decisions in relation to subsequent developments within the creator's and/or preserver's activities that might make it necessary to adjust or redo an appraisal, such as substantial changes to: (1) appraised records and/or their context, (2) the creator's organizational mandate and responsibilities, (3) the creator's record-making or recordkeeping activities or systems, (4) the preserver's records preservation activities or systems and/or (5) the preserver's organizational mandate and responsibilities

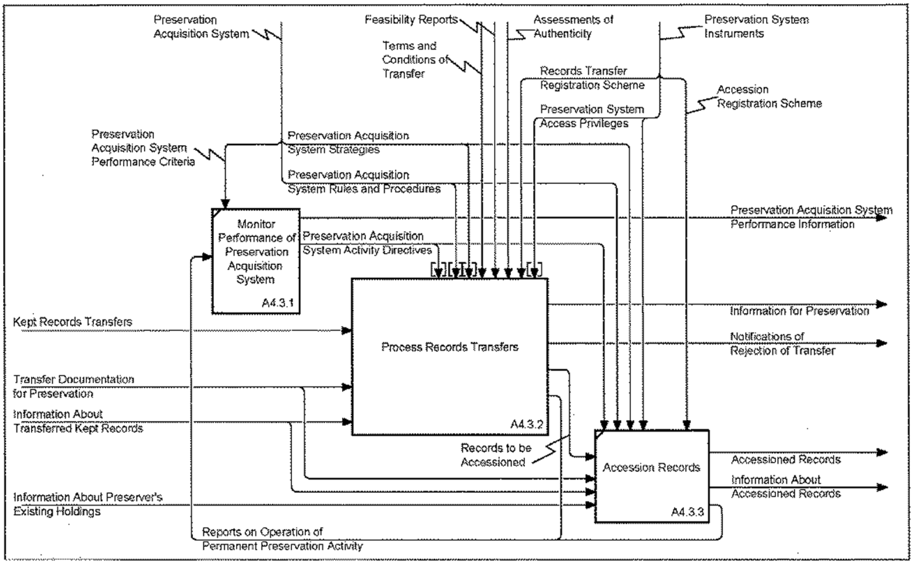

Preservation Acquisition System
Feasibility Reports
Assessments of Authenticity
Preservation System Instruments
Terms and Conditions of Transfer
Records Transfer Registration Scheme
Accession Registration Scheme
Preservation System Access Privileges
Preservation Acquisition System Performance Criteria
Preservation Acquisition System Strategies
Preservation Acquisition System Rules and Procedures
Preservation Acquisition System Performance Information
Monitor Performance of Preservation Acquisition System
A4.3.1
Preservation Acquisition System Activity Directives
Kept Records Transfers
Process Records Transfers
A4.3.2
Information for Preservation
Notifications of Rejection of Transfer
Transfer Documentation for Preservation
Information About Transferred Kept Records
Records to be Accessioned
Accession Records
A4.3.3
Accessioned Records
Information About Accessioned Records
Information About Preserver's Existing Holdings
Reports on Operation of Permanent Preservation Activity

Fig. 10. Acquiring selected records.

Table 9. Definitions for A4.3 sub-activities.

Activity	Definition
A4.3.1 Monitor performance of preservation acquisition system	To assess the efficacy of the performance of the permanent preservation acquisition system by analyzing reports on the operation of preservation activities, and issue activity directives for acquisition activities and information on the performance of the permanent preservation selection system for use in continued maintenance of the permanent preservation system
A4.3.2 Process records transfers	To register records transfers received by the designated preserver, confirm the authorization for the transfers, verify their content, confirm the authenticity of the records in the transfers and confirm the feasibility of preserving the transferred records
A4.3.2.1 Register transfers	To record registration information about received transfers and issue notifications of receipt to the persons transferring the records
A4.3.2.2 Confirm authorization for transfers	To verify the authority for transfer of records selected for preservation and, in cases of unauthorized transfers, issue notifications of rejection of transfer to the persons transferring the records
A4.3.2.3 Verify content of transfers	To determine whether transfers of records selected for preservation have been successfully transmitted (i.e., are not corrupted) and include all records and aggregates of records specified in the terms and conditions of the transfers and, in corrupted or unverified cases, issue notifications of rejection of transfer to the persons transferring the records
A4.3.2.4 Confirm authenticity of records	To determine whether the assessment of the authenticity of the creator's records being transferred, which was conducted as part of the appraisal process, is still valid by verifying that the attributes relating to the records' identity and integrity have been carried forward with them along with any relevant documentation
A4.3.2.5 Confirm feasibility of preservation	To confirm that the determinations of the feasibility of preservation made during the process of appraisal are still valid and, in unconfirmed cases, issue notifications of rejection of transfer to the persons transferring the records
A4.3.3 Accession records	To formally accept records selected for permanent preservation into custody and document transfers in accessions documentation

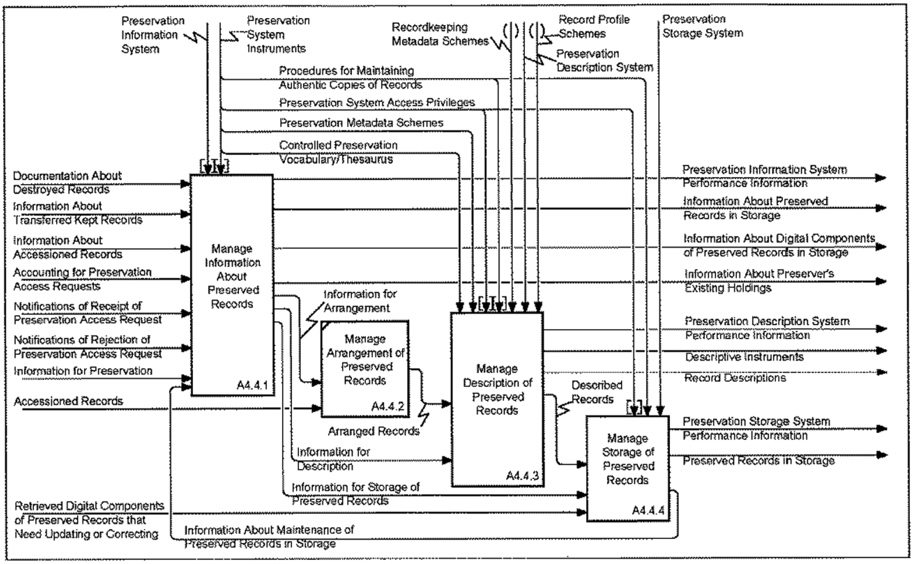

Preservation Information System
Preservation System Instruments
Recordkeeping Metadata Schemes
Record Profile Schemes
Preservation Description System
Preservation Storage System
Procedures for Maintaining Authentic Copies of Records
Preservation System Access Privileges
Preservation Metadata Schemes
Controlled Preservation Vocabulary/Thesaurus
Documentation About Destroyed Records
Information About Transferred Kept Records
Information About Accessioned Records
Accounting for Preservation Access Requests
Notifications of Receipt of Preservation Access Request
Notifications of Rejection of Preservation Access Request
Information for Preservation
Accessioned Records
Retrieved Digital Components of Preserved Records that Need Updating or Correcting
Manage Information About Preserved Records
A4.4.1
Information for Arrangement
Manage Arrangement of Preserved Records
A4.4.2
Arranged Records
Information for Description
Information for Storage of Preserved Records
Information About Maintenance of Preserved Records in Storage
Manage Description of Preserved Records
A4.4.3
Described Records
Manage Storage of Preserved Records
A4.4.4
Preservation Information System Performance Information
Information About Preserved Records in Storage
Information About Digital Components of Preserved Records in Storage
Information About Preserver's Existing Holdings
Preservation Description System Performance Information
Descriptive Instruments
Record Descriptions
Preservation Storage System Performance Information
Preserved Records in Storage

Fig. 11. Preserving accessioned records.

Table 10. Definitions for A4.4 sub-activities.

Activity	Definition
A4.4.1 Manage information about preserved records	To compile information about records in the permanent preservation system and about records preservation activities and to provide overall control and co-ordination of that information for use in records selection, acquisition, description, storage and access activities
A4.4.1.1 Monitor performance of preservation information system	To assess the efficacy of the performance of the permanent preservation information system by analyzing reports on the operation of preservation activities, and issue activity directives for information activities and information on the performance of the permanent preservation selection system for use in continued maintenance of the permanent preservation system
A4.4.1.2 Compile information for preservation	To collect, organize and record relevant appraisal, acquisition, accession and preservation information about acquired records for their preservation, description, storage, retrieval and output
A4.4.1.3 Update information on preservation actions	To record information about actions taken to back-up, correct, update and refresh digital components of records acquired for permanent preservation or their storage

(*continued*)

Table 10. (*continued*)

Activity	Definition
A4.4.2 Manage arrangement of preserved records	To provide overall control and co-ordination of records arrangement activities
A4.4.3 Manage description of preserved records	To provide overall control and co-ordination of records description activities, including monitoring the preservation description system, describing preserved records and developing description instruments
A4.4.3.1 Monitor performance of preservation description system	To assess the efficacy of the performance of the permanent preservation description system by analyzing reports on the operation of preservation activities, and issue activity directives for description activities and information on the performance of the permanent preservation selection system for use in continued maintenance of the permanent preservation system
A4.4.3.2 Describe preserved records	To record information about the nature and make-up of records acquired for permanent preservation and about their juridical-administrative, provenancial, procedural, documentary and technological contexts, as well as information about any changes they have undergone since they were first created
A4.4.3.3 Develop description instruments	To prepare tools that provide intellectual and physical control over the records in the preservation system, such as guides, inventories, indexes, repository locators and related finding aids
A4.4.4 Manage storage of preserved records	To provide overall control and co-ordination of the permanent preservation storage system and the records stored in the system by placing the records in storage, maintaining their digital components and monitoring the performance of the storage system
A4.4.4.1 Monitor performance of permanent preservation storage system	To assess the efficacy of the performance of the permanent preservation storage system by analyzing reports on the operation of preservation activities, and issue activity directives for storage activities and information on the performance of the permanent preservation selection system for use in continued maintenance of the permanent preservation system
A4.4.4.2 Place preserved records in storage	To place the digital components of preserved records and their metadata into storage in accordance with the procedures for maintaining authentic copies of records and the actions prescribed by the preservation storage system strategies, rules and procedures and activity directives

(*continued*)

Table 10. (*continued*)

Activity	Definition
A4.4.4.3 Maintain records in permanent preservation storage system	To monitor the storage of preserved records and their digital components, periodically back-up the permanent preservation storage system and, as necessary, correct problems with and update the digital components, and/or refresh storage media to ensure the records in the system remain accessible, legible and intelligible over time
A4.4.4.3.1 Monitor preserved records in storage	To keep track of the condition and maintenance requirements of preserved records—more specifically, their digital components and metadata–and the media on which they are stored in the permanent preservation storage system to identify storage that needs backing-up, digital components and metadata that need correcting or updating and media that need refreshing; and to issue reports on maintenance activities
A4.4.4.3.2 Back-up preservation storage system	To routinely make a copy of all digital content in the preservation storage system, including the operating system, the software applications and all digital objects in the system, for the purpose of recovery in the event of a disaster resulting in system failure or corruption, and record information about these back-up activities
A4.4.4.3.3 Correct problems with preserved records in storage	To take the actions prescribed by the preservation storage system strategies, rules and procedures and activity directives, in accordance with the procedures for maintaining authentic copies of records, to identify and eliminate problems in storage to ensure that the records remain accessible, legible and intelligible over time; and record information about these correction activities
A4.4.4.3.4 Update preserved records in storage	To carry out conversion actions on the digital components of preserved records in storage in accordance with the procedures for maintaining authentic copies of records and the actions prescribed by the preservation storage system strategies, rules and procedures and activity directives, to ensure the records remain accessible, legible and intelligible over time (such as by migration, standardization or transformation to persistent form), and record information about these updating activities
A4.4.4.3.5 Refresh media for preserved records in storage	To copy or transfer the digital components of preserved records in storage from one medium to another, or otherwise ensure the storage medium remains sound, in accordance with the procedures for maintaining authentic copies of records and the actions prescribed by the preservation storage system strategies, rules and procedures and activity directives, and record information about these media refreshment activities

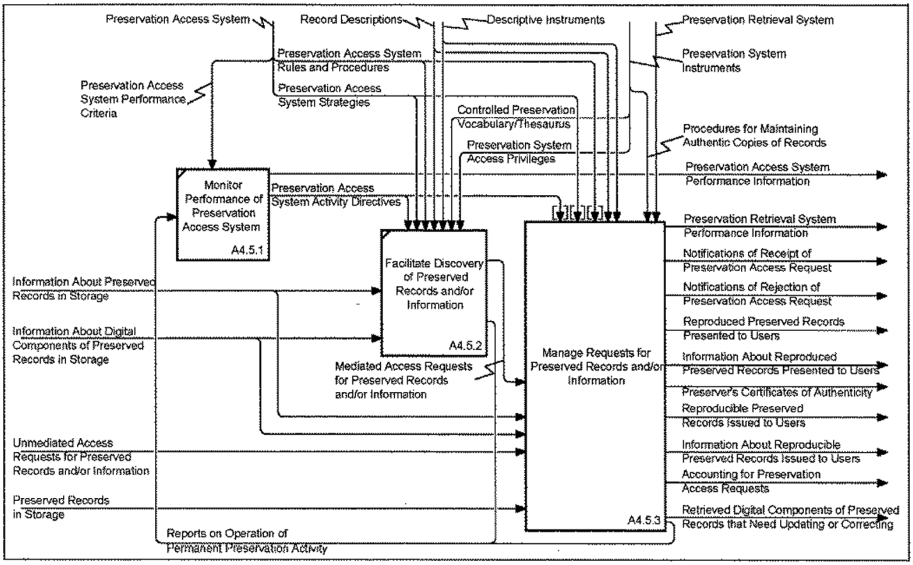

Preservation Access System
Record Descriptions
Descriptive Instruments
Preservation Retrieval System
Preservation Access System Rules and Procedures
Preservation System Instruments
Preservation Access System Performance Criteria
Preservation Access System Strategies
Controlled Preservation Vocabulary/Thesaurus
Procedures for Maintaining Authentic Copies of Records
Preservation System Access Privileges
Preservation Access System Performance Information
Monitor Performance of Preservation Access System
A4.5.1
Preservation Access System Activity Directives
Preservation Retrieval System Performance Information
Notifications of Receipt of Preservation Access Request
Facilitate Discovery of Preserved Records and/or Information
A4.5.2
Information About Preserved Records in Storage
Notifications of Rejection of Preservation Access Request
Reproduced Preserved Records Presented to Users
Information About Digital Components of Preserved Records in Storage
Manage Requests for Preserved Records and/or Information
A4.5.3
Information About Reproduced Preserved Records Presented to Users
Mediated Access Requests for Preserved Records and/or Information
Preserver's Certificates of Authenticity
Reproducible Preserved Records Issued to Users
Unmediated Access Requests for Preserved Records and/or Information
Information About Reproducible Preserved Records Issued to Users
Accounting for Preservation Access Requests
Preserved Records in Storage
Retrieved Digital Components of Preserved Records that Need Updating or Correcting
Reports on Operation of Permanent Preservation Activity

Fig. 12. Outputting records.

Table 11. Definitions for A4.5 sub-activities.

Activity	Definition
A4.5.1 Monitor performance of preservation access system	To assess the efficacy of the performance of the permanent preservation access system by analyzing reports on the operation of preservation activities, and issue activity directives for access activities and information on the performance of the permanent preservation access system for use in continued maintenance of the permanent preservation system
A4.5.2 Facilitate discovery of preserved records and/or information	To provide authorized internal and external users with mediated access to and, as necessary, assistance in the use of, the tools and resources needed to support querying and searching for information, records and/or records aggregates in the permanent preservation system
A4.5.3 Manage requests for preserved records and/or information	To provide overall control and co-ordination of internal and external requests for access to preserved records and/or information about the records by processing access requests, retrieving digital components for requested records and/or information, verifying retrieved components and information and providing access to retrieved records and/or information

(*continued*)

Table 11. (*continued*)

Activity	Definition
A4.5.3.1 Process requests for preserved records and/or information	To register access requests for preserved records and/or information, translate them, define request specifications, generate retrieval requests and account for any problems with processing access requests
A4.5.3.1.1 Register preservation access requests	To record registration information about received requests for access to preserved records and/or information about the records and issue notifications of receipt to the persons requesting the records
A4.5.3.1.2 Retrieve information to process preservation access requests	To gather the information, from description instruments and other preservation information, needed to process access requests for preserved records and/or information about records
A4.5.3.1.3 Generate preservation retrieval requests	To translate access requests for preserved records and/or information translated into requests to the permanent preservation storage and information systems for retrieval of the exact digital components and/or information required to fulfil the access requests
A4.5.3.1.4 Generate preservation requests specifications	To issue instructions to the preservation retrieval and access systems on how to fulfil requests for preserved records and/or information about the records based on analyses of the requests and processing information in relation to preservation retrieval and access systems' strategies, rules and procedures (including procedures for maintaining authentic copies of records) and access privileges
A4.5.3.2 Retrieve requested preserved records and/or information	To output copies of digital components of records, information about digital components of records, rendering information about records and/or content information about records retrieved from storage in the permanent preservation system in response to retrieval requests for components and/or information and in accordance with any request specifications
A4.5.3.3 Verify retrieved preserved records and/or information	To determine whether all components and information necessary to satisfy access requests for preserved records and/or information about the records have been received and can be processed for output and, in cases where digital components are encountered that need updating or correcting, redirect them, along with information about the problems encountered, to the maintenance function of the permanent preservation storage system for further action

(*continued*)

Table 11. (*continued*)

Activity	Definition
A4.5.3.4 Provide access to retrieved preserved records and/or information	To fulfil access requests by either reconstituting the retrieved digital components of preserved records and/or information in authentic form and presenting the manifested records or information to users, or by packaging the retrieved digital components with information about how to reconstitute and present the records and/or information with the appropriate extrinsic form and issuing the packaged materials to users, and account for the success or failure of either activity
A4.5.3.4.1 Reconstitute preserved records and/or information	To link or assemble all the verified digital components of requested preserved records and/or information about preserved records as necessary to reproduce and present the records and/or information in authentic form and, if necessary, redact information to meet privacy and/or copyright requirements
A4.5.3.4.2 Manifest preserved records and/or information	To present copies of the reconstituted requested preserved records and/or requested information about the records with the appropriate extrinsic form and with information about their relationships to one another (archival bond) and, if requested, produce a Certificate of Authenticity for the records copies
A4.5.3.4.3 Package preserved records and/or information for output	To combine the digital components of the requested preserved records and/or requested information about preserved records with information on how to reconstitute and manifest the records or information with the appropriate extrinsic form

Figure 11 [17] depicts the 4 sub-activities of A4.4: A4.4.1 Manage Information About Preserved Records, A4.4.2 Manage Arrangement of Preserved Records, A4.4.3 Manage Description of Preserved Records, and A4.4.4 Manage Storage of Preserved Records, with A4.4.1, A4.4.3, and A4.4.4 showing to contain their own sub-activities. Table 10 provides definitions for these sub-activities including those to A4.4.4.3.

Figure 12 [17] depicts the 3 sub-activities of A4.5: A4.5.1 Monitor Performance of Preservation Access System, A4.5.2 Facilitate Discovery of Preserved Records and/or Information, and A4.5.3 Manage Requests for Preserved Records and/or Information, with A4.5.3 showing to contain sub-activities. Table 11 provides definitions for these sub-activities including those to A4.5.3.1 and A4.5.3.4.

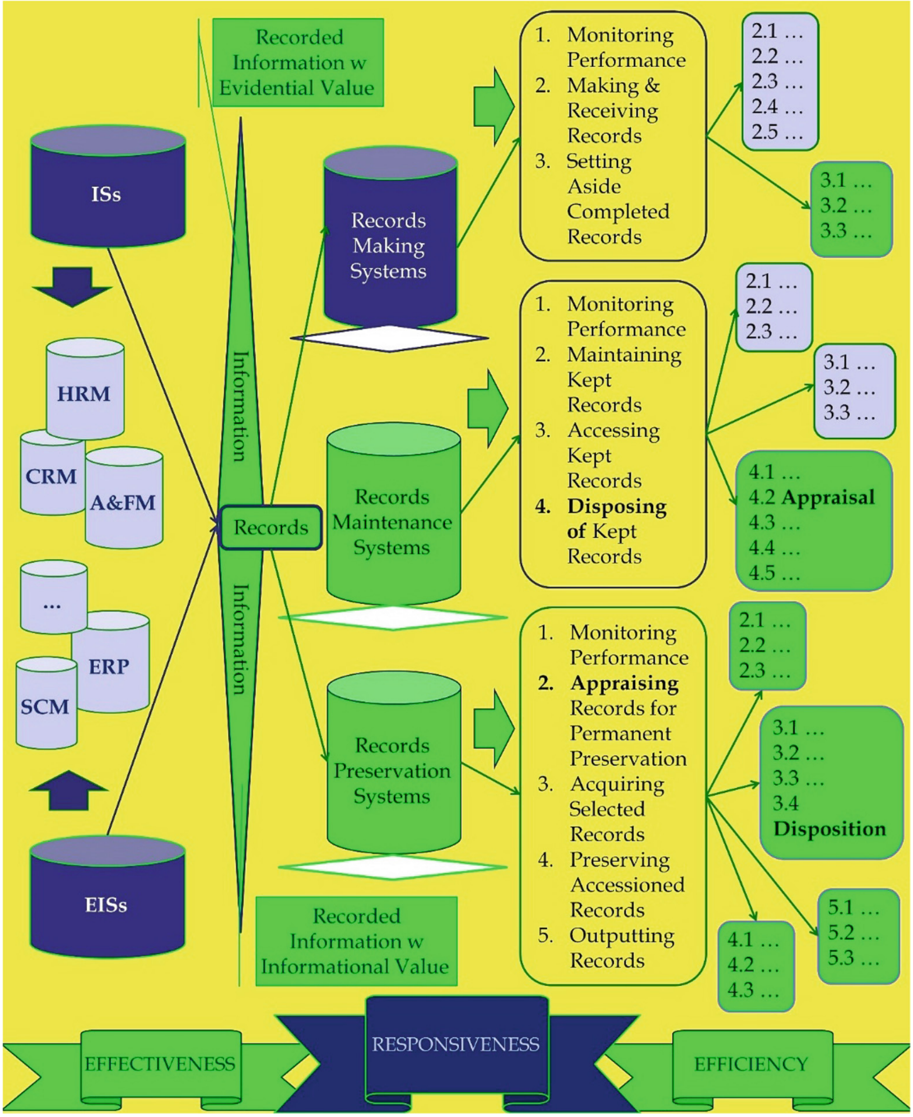

Recorded Information w Evidential Value
ISs
HRM
CRM
A&FM
...
ERP
SCM
EISs
Information
Records
Information
Records Making Systems
Records Maintenance Systems
Records Preservation Systems
1. Monitoring Performance
2. Making & Receiving Records
3. Setting Aside Completed Records
2.1 …
2.2 …
2.3 …
2.4 …
2.5 …
3.1 …
3.2 …
3.3 …
1. Monitoring Performance
2. Maintaining Kept Records
3. Accessing Kept Records
4. Disposing of Kept Records
2.1 …
2.2 …
2.3 …
3.1 …
3.2 …
3.3 …
4.1 …
4.2 Appraisal
4.3 …
4.4 …
4.5 …
1. Monitoring Performance
2. Appraising Records for Permanent Preservation
3. Acquiring Selected Records
4. Preserving Accessioned Records
5. Outputting Records
2.1 …
2.2 …
2.3 …
3.1 …
3.2 …
3.3 …
3.4 Disposition
4.1 …
4.2 …
4.3 …
5.1 …
5.2 …
5.3 …
Recorded Information w Informational Value
EFFECTIVENESS
RESPONSIVENESS
EFFICIENCY

Fig. 13. ISs/EISs and records systems.

5 Conclusion

From the above illustration, it can be concluded that for the digital records management field, ISs and EISs are record-making systems. These systems handle data and information flows to support business operations and are motivated by efficiency and convenience. Depending on the specific designs and/or clients' customization requirements, such primarily record-making systems may incorporate certain recordkeeping functionalities, yet it can hardly be said that ISs and EISs are sufficient in satisfying the requirements of organizational records management. For records to be managed as records, the control of records must go beyond individual information systems and the understanding of the value of the records must be from the viewpoint of the entire enterprise, taking into considerations of both internal and external requirements. The design and the articulation of the three types of records systems as depicted in the COP model are driven by this holistic view. ISs and EISs are designed to streamline the conduct of business activities and are equipped with standard features of control and security. They may very well be sufficient for current operation and immediate usage but inherently, they are not ready for such complicated processes as records appraisal and disposition – the hallmarks of records management. Figure 13 depicts the high-level relationships between records systems and ISs/EISs.

It is not unusual for a digital records management system to be implemented in enterprises, such as those certified by the DoD5015.2-STD or the MoReq 2010 Specification, to manage their unstructured digital records. However, transaction-oriented ISs and certainly EISs are normally left out of the control of organizational digital records management program, an issue that is not fully acknowledged by the ISs and EIS fields. ISs and EISs have been continuously advancing, and with the increasingly wide deployment of cloud infrastructure/services, they are becoming more powerful and ubiquitous. Still, these systems lack typically the functionalities that focus on systematic and consistent management of organizational information in the form of records, a stance that views the enterprise as a whole and as an integral part of society. As such, future research needs to focus on concrete cases and specific types of information systems for the purpose of establishing principles and guidelines for system designs and implementations that take into considerations of all relevant factors. It is a call by this chapter, therefore, to forge meaningful collaborations between the records profession and the ISs/EISs field so that the joint force can collectively ensure the trustworthiness of organizational digital records, maximize their value realization, and guarantee records accessibility for as not only long as the enterprise exists but also as society needs them.

Acknowledgements. The research reported in this article was supported financially by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China [Grant Number 15XNL032].

References

1. Xie, S.L.: Co-design of information systems with digital records management a proposal for research. In: Fred, A., et al. (eds.) Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2015, vol. 3, pp. 222–228. Scitepress, Setúbal (2016)
2. Xie, S.L., Fan, G.Y.: Organizational records systems - an alternative view to (enterprise) information systems. In: Fred, A., et al. (eds.) Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2015, vol. 3, pp. 82–91. Scitepress, Setúbal (2016)

3. Duranti, L.: Diplomatics: New Uses for An Old Science. SAA, ACA and Scarecrow Press, Chicago (1998)
4. Canning, R.: Electronic Data Processing for Business and Industry. Wiley, New York (1956)
5. Swanson, E.B.: Information systems. In: Bates, M.J., Maack, M.N. (eds.) Encyclopedia of Library and Information Sciences, 3rd edn, pp. 2635–2642. Taylor and Francis, New York (2009)
6. Date, C.J.: An Introduction to Data Base Systems, 3rd edn. Addison-Wesley, Reading (1981)
7. International Organization for Standardization: ISO 15489-1 Information and documentation–records management. Part 1: General (2001)
8. National Archives of Australia: A-Z for information and records management. http://naa.gov.au/information-management/support/a-z/index.aspx. Accessed 23 Apr 2018
9. OMB: Management of federal information resources. https://www.whitehouse.gov/omb/circulars_a130_a130trans4/#3. Accessed 23 Apr 2018
10. TBS: Policy on information management. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=12742. Accessed 23 Apr 2018
11. TBS: Directive on recordkeeping. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=16552. Accessed 23 Apr 2018
12. U.S. 44 U.S.C: Chapter 33 disposal of records. http://uscode.house.gov/view.xhtml?req=(title:44%20section:3301%20edition:prelim). Accessed 23 Apr 2018
13. Hirschheim, R., Klein, H.K.: Tracing the history of the information systems field. In: Galliers, R.D., Currie, W.L. (eds.) The Oxford Handbook of Management Information Systems: Critical Perspectives and New Directions, pp. 16–61. Oxford University Press, Oxford (2011)
14. Xu, X.: Editorial: inaugural issue. Enterp. Inf. Syst. **1**(1), 1–2 (2007)
15. Parthasarathy, S.: Enterprise Information Systems and Implementing IT Infrastructures: Challenges and Issues. Hershey, New York (2010)
16. U.S. 44 U.S.C: Chapter 29 Records Management by the Archivist of the United States. https://www.law.cornell.edu/uscode/text/44/chapter-29. Accessed 23 Apr 2018
17. Eastwood, T., et al.: Appendix 14: Chain of preservation model – diagrams and definitions. In: Duranti, L., Preston, R. (eds.) InterPARES 2: Experiential, Interactive and Dynamic Records (2008). http://www.interpares.org/display_file.cfm?doc=ip2_book_appendix_14.pdf. Accessed 23 Apr 2018

Correction to: Knowledge Discovery, Knowledge Engineering and Knowledge Management

Ana Fred, Jan Dietz, David Aveiro, Kecheng Liu, Jorge Bernardino, and Joaquim Filipe

Correction to:
A. Fred et al. (Eds.): *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, CCIS 914, https://doi.org/10.1007/978-3-319-99701-8

In a previous version of this publication, the affiliation of the second editor was incomplete. This has now been corrected.

In a previous version of this publication, the family name of the first author in the paper titled "The Mereologies of Upper Ontologies" was incorrect. This has now been corrected.

The updated version of the book can be found at
https://doi.org/10.1007/978-3-319-99701-8
https://doi.org/10.1007/978-3-319-99701-8_9

A. Fred et al. (Eds.): IC3K 2016, CCIS 914, p. C1, 2019.
https://doi.org/10.1007/978-3-319-99701-8_19

Author Index