ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING 183

Aleksander Zgrzywa
Kazimierz Choroś
Andrzej Siemiński (Eds.)

Multimedia and Internet Systems: Theory and Practice

Springer

Advances in Intelligent Systems and Computing 183

For further volumes:
http://www.springer.com/series/11156

Aleksander Zgrzywa, Kazimierz Choroś,
and Andrzej Siemiński (Eds.)

Multimedia and Internet Systems: Theory and Practice

Springer

Editors
Aleksander Zgrzywa
Institute of Informatics
Wrocław University of Technology
Wrocław
Poland

Andrzej Siemiński
Institute of Informatics
Wrocław University of Technology
Wrocław
Poland

Kazimierz Choroś
Institute of Informatics
Wrocław University of Technology
Wrocław
Poland

ISSN 2194-5357
e-ISSN 2194-5365
ISBN 978-3-642-32334-8
e-ISBN 978-3-642-32335-5
DOI 10.1007/978-3-642-32335-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012943389

Printed on acid-free paper

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

Preface

During the last 20 years we have witnessed a rapid development of Multimedia and Network Information Systems. What is even more important, the pace of change does not show any sign of slowing. When we look back we see how many research projects that have originated at universities or in research facilities now are part of our everyday life. This calls for a volume that addresses the capabilities, limitations, current trends and perspectives of Multimedia and Network Information Systems.

Our intention is to offer the readers of this monograph a very broad review of the recent scientific problems in that area. Solving them has became a principal task of numerous scientific teams all over the world. The volume is a selection of representative investigations, solutions and applications submitted by scientific teams working in many European countries.

Content of the book has been divided into four parts:

I Multimedia Information Technology
II Information System Specification
III Information System Applications
IV Web Systems and Network Technologies

Part I contains eight chapters that discusses new methods of visual data processing. The studies and resulting solutions described in the several chapters of this part follow the gaining momentum trend of exploiting artificial intelligence techniques, fuzzy logic, multi-agent approaches in the domain of multimedia information technology. The domain covers image alignment, video deinterlacing, cartographic representation, visual objects description, large scale 3D reconstruction, and hand gesture recognition. Two of the chapters focus on the content-based indexing and retrieval of visual data.

The second part of the book consists of seven chapters. Two of them address the specific problems of different applications of ontology's in information systems and the ontology alignment. This part also contains reports on the experiments on the evaluation of re-sampling combined with clustering and random oracle using genetic fuzzy systems and the study of parameter selection for the Dynamic Travelling Salesman Problem. One of the chapters deals with the modeling failures of distributed systems in stochastic process algebras. The important from a practical point of view problem of tracking changes in database schemas is also discussed in the last chapter of this part.

Part III presents a handful of applications. It consists of four chapters. One of them describes an adaptive e-learning application with a catchy name of Power Chalk. Innovative solutions to the problem of unified user identification for mobile environments and smart communications for remote users are also discussed. The last chapter presents a rule based expert system that covers semantic matching, spatiotemporal relation operators, and comparison of GIS data to eliminate VAT-carousel crimes.

Part IV refers to the trends and perspectives in Web Systems and Network Technologies. It contains 6 chapters. Two of them deal with e-commerce issues discussing the evaluation of the server performance and presenting an approach to use product reviews from e-commerce websites for the product feature opinion mining task. The usefulness of Web pages can considerably suffer from poor readability. Therefore a special chapter is devoted to a methodology of creating a computer aided system for measuring text readability. Finding relevant services from a service collection is yet another important aspect of the Web Technologies. Two of the chapters in this part are devoted to the problem.

The book should be of great interest to researchers involved in all aspects of multimedia and Internet applications. We hope that it will fulfill the expectations of its readers and we will be also very pleased if the book will attract more scholars to work on the area and to inspire the research community already working on the domain. If so, the goal that motivated authors, reviewers, and editors will be achieved. It will be also the greatest prize for our joint efforts.

Aleksander Zgrzywa
Kazimierz Choroś
Andrzej Siemiński

Contents

Part I: Multimedia Information Technology

Part II: Information Systems Specification

Part III: Information Systems Applications

List of Contributors

Jarosław Bąk
Poznań University of Technology
Institute of Control and
Information Engineering
pl. M. Skłodowskiej-Curie 5
60-965 Poznań
Poland
`jaroslaw.bak@put.poznan.pl`

Radosław Bednarski
Lodz University of Technology
Institute of Information Technology
ul. Żeromskiego 116
90-924 Łódź
Poland
`radoslaw.bednarski@p.lodz.pl`

Miroslav Behan
University of Hradec Kralove
FIM, Department of Information
Technologies
Rokitanskeho 62, Hradec Kralove
50003, Czech Republic
`miroslav.behan@uhk.cz`

Leszek Borzemski
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław
Poland
`leszek.borzemski@pwr.wroc.pl`

Peter Butka
Technical University of Košice
Faculty of Electrical Engineering
and Informatics
Department of Cybernetics and
Artifcial Intelligence
Letńa 9, 04200 Košice
Slovakia
`peter.butka@tuke.sk`

Jerzy Brzeziński
Poznań Univerity of Technology
Institute of Computing Science
Piotrowo 2, 60-965, Poznań
Poland
`jerzy.brzezinski@`
`put.poznan.pl`

Kazimierz Choroś
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław, Poland
`kazimierz.choros@pwr.wroc.pl`

Adam Czyszczoń
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław, Poland
`adam.czyszczon@pwr.wroc.pl`

Andrzej Czyżewski
Gdańsk University of Technology
ul. Narutowicza 11/12
80-233 Gdańsk
Poland
andcz@sound.eti.pg.gda.pl

Dariusz Dwornikowski
Poznań Univerity of Technology
Institute of Computing Science
Piotrowo 2, 60-965, Poznań
Poland
dariusz.dwornikowski@
cs.put.poznan.pl

Damian Ellwart
Gdańsk University of Technology
Narutowicza 11/12
80-233 Gdańsk
Poland
ellwart@sound.eti.pg.gda.pl

Margarita Esponda-Argüero
Freie Universität Berlin
Department of Mathematics
and Computer Science
Takustrasse 9, 14195 Berlin
Germany
esponda@inf.fu-berlin.de

Maciej Falkowski
Poznań University of Technology
Institute of Control and Information
Engineering
pl. M. Skłodowskiej-Curie 5
60-965 Poznań
Poland
maciej
falkowski@put.poznan.pl

Piotr Gawrysiak
Warsaw University of Technology
Institute of Computer Science
ul. Nowowiejska 15/19, 00-665
Warszawa
Poland
p.gawrysiak@ii.pw.edu.pl

Tatiana Jaworska
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warszawa
Poland
tatiana.jaworska@
ibspan.waw.pl

Czesław Jędrzejek
Poznań University of Technology
Institute of Control and Information
Engineering
pl. M. Skłodowskiej-Curie 5
60-965 Poznań, Poland
czeslaw.jedrzejek@
put.poznan.pl

Marek Kopel
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław, Poland
marek.kopel@pwr.wroc.pl

Bożena Kostek
Multimedia Systems Department
Gdańsk University of Technology
ul. Narutowicza 11/12
80-233 Gdańsk, Poland
bozenka@sound.eti.pg.gda.pl

Ondřej Krejcar
University of Hradec Kralove
FIM, Department of Information
Technologies
Rokitanskeho 62, Hradec Kralove,
50003
Czech Republic
ondrej.krejcar@asjournal.eu

Tadeusz Lasota
Wrocław University of Environmental
and Life Sciences
Department of Spatial Management
ul. Norwida 25/27
50-375 Wrocław
Poland
tadeusz.lasota@up.wroc.pl

Michał Lech
Gdańsk University of Technology
Multimedia Systems Department
ul. Narutowicza 11/12, 80-233
Gdańsk
Poland
mlech@sound.eti.pg.gda.pl

Florian Liefers
Tuebingen University, Germany
Geschwister-Scholl-Platz
72074 Tübingen
forian@liefers.com

Karol Lisowski
Gdańsk University of Technology
Multimedia Systems Department
ul. Narutowicza 11/12
80-233 Gdańsk
Poland
lisowski@sound.eti.pg.gda.pl

Tomáš Machálek
University of Hradec Králové
Faculty of Informatics and Management
Department of Information
Technologies, Rokitanského 62
Hradec Králové, 50003
Czech Republic
tomas.machalek@uhk.cz

Jakub Marciniak
Adam Mickiewicz University
Faculty of Mathematics
and Computer Science
ul. Umultowska 87
61-614 Poznań
Poland
kubam@amu.edu.pl

Zygmunt Mazur
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław
Poland
Zygmunt.Mazur@pwr.wroc.pl

Ngoc Thanh Nguyen
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław
Poland
ngoc-thanh.nguyen@pwr.wroc.pl

Maciej Nowak
Poznań University of Technology
Institute of Control and Information
Engineering
pl. M. Skłodowskiej-Curie 5
60-965 Poznań
Poland
maciej.nowak@put.poznan.pl

Kamila Olševičová
University of Hradec Králové
Faculty of Informatics and Management
Department of Information
Technologies Rokitanského 62
Hradec Králové, 500 03, Czech
Republic
kamila.olsevicova@uhk.cz

Tadeusz Pankowski
Poznań University of Technology
Institute of Control and Information
Engineering
pl. M. Skłodowskiej-Curie 5
60-965 Poznań, Poland
tadeusz.pankowski@
put.poznan.pl

Roman Parys
Tuebingen University, Germany
Geschwister-Scholl-Platz
72074 Tübingen
parys@gris.uni-tuebingen.de

Marcin Pietranik
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław
Poland
marcin.pietranik@pwr.wroc.pl

Maria Pietruszka
Lodz University of Technology
Institute of Information Technology
ul. Żeromskiego 116, 90-924 Łódź
Poland
maria.pietruszka@p.lodz.pl

Jozef Pócs
Slovak Academy of Sciences
Mathematical Institute
Grešakova 6, 040 01 Košice, Slovakia
pocs@saske.sk

Jana Pócsová
Technical University of Košice
BERG Faculty, Institute of Control
and Informatization of Production
Processes, Boženy Němcovej 3
043 84 Košice, Slovakia
jana.pocsova@tuke.sk

Raúl Rojas
Freie Universität Berlin
Department of Mathematics
and Computer Science
Takustrasse 9, 14195 Berlin, Germany
raul.rojas@fu-berlin.de

Dan-El Neil Vila Rosado
Freie Universität Berlin
Department of Mathematics
and Computer Science
Takustrasse 9, 14195 Berlin, Germany
vila80@inf.fu-berlin.de

Martin Sarnovský
Technical University of Košice
Faculty of Electrical Engineering and
Informatics
Department of Cybernetics and Artifcial
Intelligence
Letňa 9, 04200 Košice, Slovakia
martin.sarnovsky@tuke.sk

Andreas Schilling
Tuebingen University, Germany
Geschwister-Scholl-Platz
72074 Tübingen
schilling@uni-tuebingen.de

Andrzej Siemiński
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław
Poland
andrzej.sieminski@
pwr.wroc.pl

Janusz Sobecki
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław
Poland
Janusz.Sobecki@pwr.wroc.pl

Grażyna Suchacka
Opole University, Faculty of
Mathematics
Physics and Computer Science
ul. Oleska 48
45-052 Opole
Poland
g.suchacka@gmail.com

Zbigniew Telec
Wrocław University of Technology
Institute of Informatics, Wybrzeże
Wyspiańskiego 27
50-370 Wrocław
Poland
zbigniew.telec@pwr.wroc.pl

Bogdan Trawiński
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław, Poland
bogdan.trawinski@pwr.wroc.pl

Grzegorz Trawiński
Wrocław University of Technology
Faculty of Electronics
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
grzegorztrawinski@wp.pl

Maria Trocan
Institut Supérieur d'Electronique
de Paris Signal
and Image Processing Department
28 rue Notre Dame des Champs
Paris, France
maria.trocan@isep.fr

Bartomiej Twardowski
Warsaw University of Technology
Institute of Computer Science
ul. Nowowiejska 15/19
00-665 Warszawa
Poland
b.twardowski@ii.pw.edu.pl

Konrad Wiklak
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
konrad.wiklak@gmail.com

Aleksander Zgrzywa
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370, Wrocław
Poland
aleksander.zgrzywa@
pwr.wroc.pl

Part I
Multimedia Information Technology

Chapter 1
Fuzzy Rule-Based Classifier for Content-Based Image Retrieval

Tatiana Jaworska

Abstract. At present a great deal of research is being done in different aspects of Content-Based Image Retrieval System (CBIR)**.** Thus, it is necessary to develop appropriate information systems to efficiently manage datasets. Image classification is one of the most important services in image retrieval that must support these systems. The primary issue we have addressed is: how can the fuzzy set theory be used to handle crisp data for images. We propose how to introduce fuzzy rule-based classification for image objects. To achieve this goal we have constructed fuzzy rule-based classifiers, taking into account crisp data. In this chapter we present the results of the use of this fuzzy rule-based system in our CBIR.

1.1 Introduction

In recent years, the availability of image resources on the WWW and large image datasets has increased tremendously. This has created a demand for effective and flexible techniques for automatic image classification and retrieval. Although attempts to perform the Content-Based Image Retrieval (CBIR) in an efficient way have been made before, a major problem in this area has been computer perception, to which it is hard to introduce an additional semantic data model. In other words, there is a necessity to introduce fuzzy information models into image retrieval, based on high-level semantic concepts that perceive an image as a complex whole.

Images and graphical data are complex in terms of visual and semantic contents. Depending on the application, images are modelled and indexed using their

- visual properties (or a set of relevant visual features),
- semantic properties,
- spatial or temporal relationships of graphical objects.

Tatiana Jaworska
Polish Academy of Sciences, Systems Research Institute, Warsaw, Poland
e-mail: `Tatiana.Jaworska@ibspan.waw.pl`

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 3–13.
springerlink.com

Over the last decade a number of concepts of the CBIR [1], [2], [3], [4], have been used. Proposals can be found for the relational [5], object-oriented [6], [7] databases. For about 10 years the fuzzy proposition has been applied in object-relation database models [8], [9]. Zadeh's fuzzy sets theory has allowed us to develop limited programming tools, concerned with graphical applications and dealing with imperfect pictorial data. Within the scope of semantic properties, as well as graphical object properties, the first successful attempt was made by Candan and Li [10], who constructed the Semantic and Cognition-based Image Retrieval (SEMCOG) query processor to search for images by predicting their semantic and spatial imperfection.

The feature vector is used for tentative object classification at the local level of a separated object. First, we have to classify objects in order to assign them to a particular class which is later used to describe spatial relationships characteristic of a particular image. In our system spatial object location in an image is used as the global feature; then it supports full identification of graphical elements based on rules of location. Next, classified objects are used to enable the user to compose their own image in the GUI. Finally, we apply classes in order to compare objects coming from different images [21].

We have chosen a fuzzy rule-based classification to support our pattern library which is constructed to enable the user to build their image query in as natural a way as possible. 'Natural' here means handling such objects as houses, trees, water instead of a red square, blue rectangle, etc.

In this chapter we present the fuzzy rule-based classifiers for object classification which takes into account object features, together with different spatial location of segmented objects in the image. In order to improve the comparison of two images, we need to label these objects in a semantic way.

In general, our system consists of four main blocks:

1. the image preprocessing block (responsible for image segmentation), applied in Matlab, cf. [11];
2. the Oracle Database, storing information about whole images, their segments (here referred to as graphical objects), segment attributes, object location, pattern types and object identification, cf. [12];
3. the search engine responsible for the searching procedure and retrieval process based on feature vectors for objects and spatial relationship of these objects in an image, applied in Matlab;
4. the graphical user's interface (GUI) allows users to compose their own image, consisting of separate graphical objects as a query. Classification helps in the transition from rough graphical objects to human semantic elements. We have had to create a user-friendly semantic system, also applied in Matlab.

There have been several attempts to design efficient, invariant, flexible and intelligent image archival and retrieval systems based on the perception of spatial relationships. Chang [13] proposed a symbolic indexing approach, called the nine directional lower triangular (9DLT) matrix to encode symbolic images. Using the concept of 9DLT matrix, Chang and Wu [14] proposed an exact match of the retrieval scheme, based upon principal component analysis (PCA). Unfortunately, it transpired that the first principal component vectors (PCVs) associated with the image and the same image rotated are not the same. Eventually, an invariant scheme for retrieval of symbolic images based upon the PCA was prepared by Guru and Punitha [15].

1.2 Fuzziness

1.2.1 Fuzzy Sets

Let $T = \{t_1,\ldots,t_n\}$ be a set of objects, where $t_i \neq t_j$, the attributes of t_i and t_j are different. Let $U = \{A_1, \ldots, A_m\}$ be the set of attributes over $t_1,\ldots,t_n$. Each attribute A_i has been associated with a domain, denoted by $U(A_i)$. The value of attribute A_i is either a crisp or fuzzy value. According to Zadeh [16], a fuzzy set F in $U \subseteq \mathbb{R}$ is uniquely specified by its membership function μ_i: $U \rightarrow [0,1]$. Thus, the fuzzy set is described as follows:

$$F = \{(u, \mu_F(u)) | u \in U\} \tag{1.1}$$

Two important concepts of *core* and *support* are related to a fuzzy set F:

$$core\ (F) = \{u \mid u \in U \wedge \mu_F(u) = 1\}$$

and

$$support\ (F) = \{u \mid u \in U \wedge \mu_F(u) > 0\}.$$

For our purpose, we use a trapezoidal membership function MF which is mathematically defined by four parameters $\{a,b,c,d\}$:

$$\mathbf{trap\,mf}(u; a,b,c,d) = \begin{cases} 0, & u < a \\ \dfrac{u-a}{b-a}, & a \leq u \leq b \\ 1, & b \leq u \leq c \\ \dfrac{d-u}{d-c}, & c \leq u \leq d \\ 0, & d < u \end{cases} \tag{1.2}$$

Let F and G be two fuzzy sets in the universe U, we say that $F \subseteq G \Leftrightarrow \mu_F(u) \leq \mu_G(u)$, $\forall\ u \in U$. The complement of F, denoted by F^c, is defined by $\mu_{F^c}(u) = 1 - \mu_F(u)$. Furthermore, $F \cap G$ (respectively $F \cup G$) is defined the following way: $\mu_{F \cap G} = \min(\mu_F(u), \mu_G(u))$ (respectively $\mu_{F \cup G} = \max(\mu_F(u), \mu_G(u))$).

1.2.2 Fuzzy Rule-Based Classifiers

We assume that we have an M-class pattern classification problem in an n-dimensional normalized hyper-cube $[0, 1]^n$. For this problem, we use fuzzy rules of the following type:

$$\text{Rule } R_q : \text{If } x_1 \text{ is } A_{q1} \text{ and ... and } x_n \text{ is } A_{qn} \text{ then Class } C_q \text{ with } CF_q, \tag{1.3}$$

where R_q is the label of the q^{th} fuzzy rule, $\mathbf{x} = (x_1, \ldots, x_n)$ is an n-dimensional pattern/feature vector, A_{qi} is an antecedent fuzzy set ($i = 1,\ldots,n$), C_q is a class label,

CF_q is a real number in the unit interval [0,1] which represents a rule weight. The rule weight can be specified by a heuristic manner or it can be adjusted by a learning algorithm [17], [18]. We use the n-dimensional fuzzy vector $\boldsymbol{A}_q = (A_{q1}, ..., A_{qn})$ to represent the antecedent part of the fuzzy rule R_q in (3) in a concise manner.

Let S be a set of fuzzy rules of the type in (3). That is, S is a fuzzy rule-based classifier. When an n-dimensional pattern/feature $\mathbf{x}_p = (x_{p1}, ..., x_{pn})$ is presented to S, first the compatibility grade of $\mathbf{x}_p$ with the antecedent part $\boldsymbol{A}_q$ of each fuzzy rule R_q in S is calculated by the product operator as

$$\mu_{\mathbf{A}q}(\mathbf{x}_p) = \mu_{\mathbf{A}q1}(\mathbf{x}_{p1}) \times ... \times \mu_{\mathbf{A}qn}(\mathbf{x}_{pn}) \text{ for } R_q \in S, \tag{1.4}$$

where $\mu_{\mathbf{A}qi}(.)$ shows the membership function of A_{qi}. Then a single winner rule $R_{w(\mathbf{x}^p)}$ is identified for $\mathbf{x}_p$ as follows:

$$CF_{w(\mathbf{x}^p)} \times \mu_{\mathbf{A}_{w(\mathbf{x}^p)}}(\mathbf{x}_p) = \max \{CF_q \times \mu_{\mathbf{A}q}(\mathbf{x}_p) \mid R_q \in S\}, \tag{1.5}$$

where $w(\mathbf{x}_p)$ denotes the rule index of the winner rule for $\mathbf{x}_p$.

The pattern $\mathbf{x}_p$ is classified by the single winner rule $R_{w(\mathbf{x}^p)}$ as its consequent class. If there is no fuzzy rule with a positive compatibility grade with $\mathbf{x}_p$ (i.e., if $\mathbf{x}_p$ is not covered by any fuzzy rules in S), the classification of $\mathbf{x}_p$ is rejected. The classification of $\mathbf{x}_p$ is also rejected if multiple fuzzy rules with different consequent classes have the same maximum value on the right-hand side of (5). In this case, $\mathbf{x}_p$ is exactly on the classification boundary between the different classes.

We use the single winner-based fuzzy reasoning method in (5) for pattern classification. This is because a responsible fuzzy rule can be always identified for the classification result of each input pattern when we use the single winner-based fuzzy reasoning method.

An ideal theoretical example of a simple three-class, two-dimensional pattern classification problem with 20 patterns from each class is considered in [19]. There three linguistic values (*small*, *medium* and *large*) were used as antecedent fuzzy sets for each of the two attributes and 3×3 fuzzy rules were generated. S_1 was the fuzzy rule-based classifier with the 9 fuzzy rules shown below:

S_1: fuzzy rule-based classifier with 9 fuzzy rules

R_1: If x_1 is *small* and x_2 is *small* then Class2 with 1.0,
R_2: If x_1 is *small* and x_2 is *medium* then Class2 with 1.0,
R_3: If x_1 is *small* and x_2 is *large* then Class1 with 1.0,
R_4: If x_1 is *medium* and x_2 is *small* then Class2 with 1.0,
R_5: If x_1 is *medium* and x_2 is *medium* then Class2 with 1.0,
R_6: If x_1 is *medium* and x_2 is *large* then Class1 with 1.0,
R_7: If x_1 is *large* and x_2 is *small* then Class3 with 1.0,
R_8: If x_1 is *large* and x_2 is *medium* then Class3 with 1.0,
R_9: If x_1 is *large* and x_2 is *large* then Class3 with 1.0,

R_3	R_6	R_9
R_2	R_5	R_8
R_1	R_4	R_7

Fig. 1.1 Classification boundaries for fuzzy rule-based classifier S1.

For simplicity, the rule weight is 1.0 in S_1. The location of each rule is shown in Fig. 1.1.

1.3 Graphical Data Representation

In our system, Internet images are downloaded. Firstly, the new image is segmented, creating a collection of objects. Each object, selected according to the algorithm presented in detail in [21], is described by some low-level features. The features describing each object include: average colour k_{av}, texture parameters T_p, area A, convex area A_c, filled area A_f, centroid $\{x_c, y_c\}$, eccentricity e, orientation α, moments of inertia m_{11}, bounding box $\{bb_1(x,y), ..., bb_s(x,y)\}$ (s – number of vertices), major axis length m_{long}, minor axis length m_{short}, solidity s and Euler number E and Zernike moments $Z_{00},\ldots,Z_{33}$. All features, as well as extracted images of graphical objects, are stored in the DB. Let Fo be a set of features where:

$$F_O = \{k_{av}, T_p, A, A_c, ..., E\} \tag{1.6}$$

For ease of notation we will use $F_O = \{f_1, f_2, \ldots, f_r\}$, where r – number of attributes. For an object, we construct a feature vector O containing the above-mentioned features:

$$O = \begin{bmatrix} O(k_{av}) \\ O(T_p) \\ O(A) \\ \vdots \\ O(Z_{33}) \end{bmatrix} = \begin{bmatrix} O(f_1) \\ O(f_2) \\ O(f_3) \\ \vdots \\ O(f_r) \end{bmatrix}. \tag{1.7}$$

The next complex feature attributed to objects is texture. Texture parameters are found in the wavelet domain (the Haar wavelets are used). The algorithm details are also given in [21]. The use of this algorithm results in obtaining two ranges for the horizontal object dimension h and two others for the vertical one v:

$$T_p = \begin{Bmatrix} h_{\min_{1,2}}; h_{\max_{1,2}} \\ v_{\min_{1,2}}; v_{\max_{1,2}} \end{Bmatrix}. \tag{1.8}$$

Additional features of the low level for objects are shape descriptors. They are also included in the above-mentioned feature vector. We apply the two most important shape descriptors such as moments of inertia:

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y), \qquad p,q = 0,1,2 \tag{1.9}$$

and Zernike moments [20]. Zernike moments are a set of complex polynomials $\{\mathrm{V_{pq}(x,y)}\}$ which form a complete orthogonal set over the unit disk of $x^2 + y^2 \le 1$. Hence, the definition of 2D Zernike moments with p^{th} order with repetition q for intensity function *f(x,y)* of the image is described as:

$$Z_{pq} = \frac{p+1}{\pi} \iint_{x^2+y^2\le 1} V^*_{pq}(x,y) f(x,y)\, dxdy \tag{1.10}$$

where: $$V^{*}_{pq}(x, y) = V_{p,-q}(x, y) \,. \tag{1.11}$$

For our purpose, the first 10 Zernike moments are sufficient, which means we calculate moments from Z_{00} to Z_{33}.

1.4 Classification Results

The feature vector **O** (cf. (7)) is used here as a pattern/feature vector **x** for object classification. We collected $n = 32$ features for each graphical object. Based on the data collected in our CBIR system, we have analysed the most distinguished features to present our experimental results. We have chosen three classes from graphical objects in the training subset, namely: class1 - horizontal line, class2 - caret and class3 - vertical line, presented respectively in Fig. 1.2.

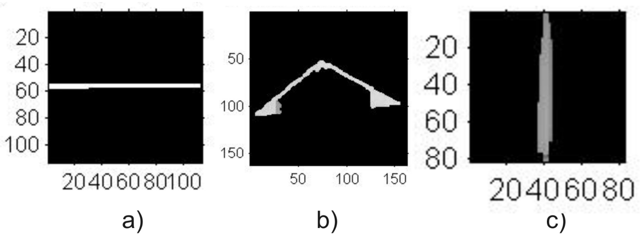


Fig. 1.2 Examples of graphical objects used as class1 - horizontal line a), class2 – caret b) and class3 - vertical line c) from the training subset.

For our fuzzy rule-based classifier we have chosen a trapezoidal MF (cf. (**1.2**)), as it better represents the character of our data. We have classified data from a training subset according to the fuzzy rule-based classifier S_1. As we mentioned earlier, in our experiment we used a three-class problem for two features: x_1 – area and x_2 – minor axis length (shown in Fig. 1.3).

Thanks to the use of the fuzzy rule-based classifier S_1, we can classify a new object (depicted as a magenta square in Fig. 1.4) from unknown class? to class2. After a comparison with the real image object, we can conclude that the classified object, in fact, belongs to class2. This confirms that we can use the single winner-based fuzzy reasoning method for our pattern classification (see Fig. 1.4).

In a multi-objective fuzzy rule-based classifier design, the accuracy of classifiers is not viewed as a factor related to interpretability. This is because accuracy is handled as a separate goal from interpretability in a multi-objective fuzzy rule-based classifier design. However, the accuracy of the winner rule seems to be an important factor, related to the explanation capacity for fuzzy rule-based classifiers [19].

Now, we show the use of a fuzzy rule-based classifier with three rules for our three-class problem, where:

S_2: fuzzy rule-based classifier with three fuzzy rules [19]
R_{123}: If x_1 is *small* then Class2 with 1.0,
R_{456}: If x_1 is *medium* then Class2 with 1.0,
R_{789}: If x_1 is *large* then Class3 with 1.0,

For this purpose we have chosen the fuzzy rule-based classifier S_2. As we have mentioned earlier, in our second experiment we used a three-class problem for two features: x_1 – minor axis length and x_2 – Zernike moment Z_{00}. We used a trapezoidal MF because the data are not normalised to the interval [0,1], according to the assumption from the fuzzy rules definition. We use the same classes (class1 - horizontal line, class2 - caret and class3 - vertical line). As it is shown in Fig. 1.5, the three rules are enough to separate the objects described by real data.

Through the use of the fuzzy rule-based classifier S_2, we can classify a new object (depicted as a magenta square in Fig. 1.6) from unknown class? to class2. After a comparison with the real image object, we can conclude that the classified object in fact belongs to class2. This confirms that we can use a single winner-based fuzzy reasoning method for our pattern classification (see Fig. 1.6).

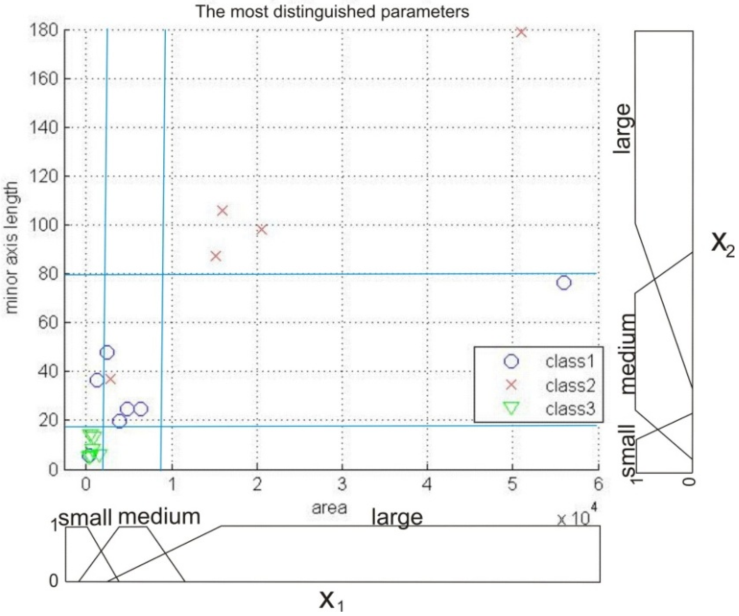


Fig. 1.3 Three-class problem for two features: x1 - minor axis length and x2 - area.

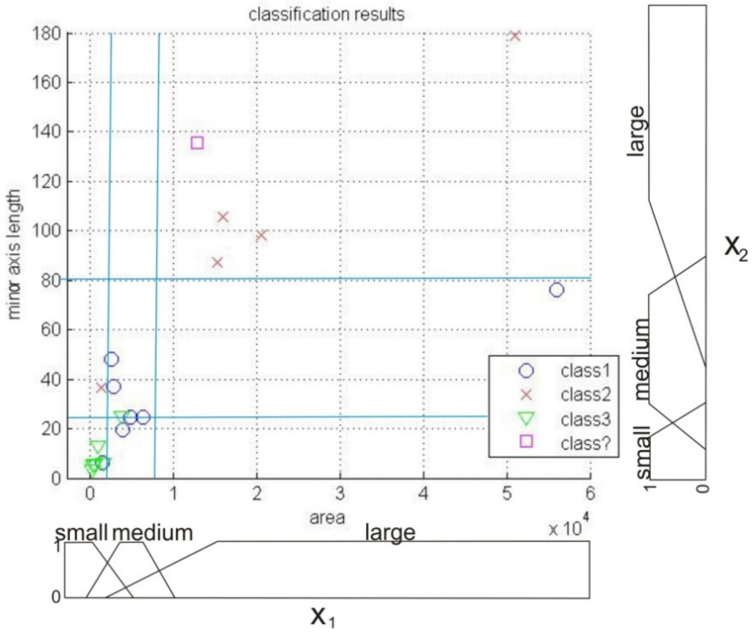


Fig. 1.4 The magenta square is a classified element for the fuzzy rule classifier S1 with 9 rules.

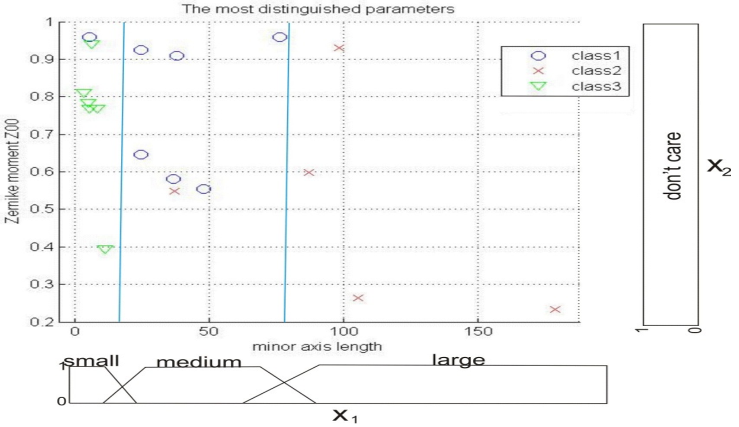


Fig. 1.5 Classification with three fuzzy rules S2

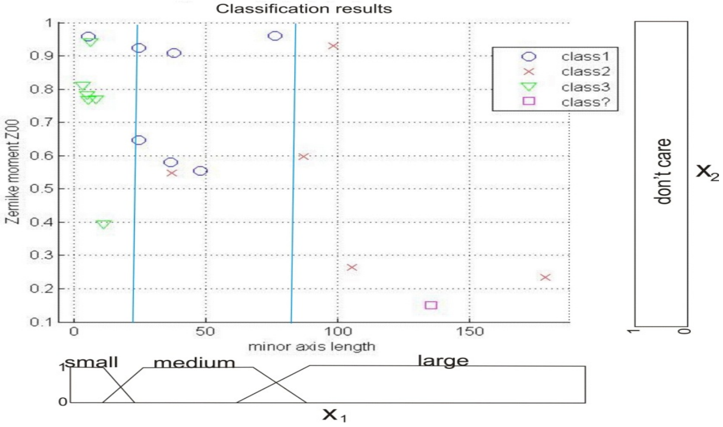


Fig. 1.6 Classification results for a classifier with three fuzzy rules S2

1.5 Further Use of Classified Objects in CBIR

Therefore, we have to classify objects in order to [21]:

1. use particular classes as patterns. We store these data in DB to use them in CBIR algorithms.
2. specify a spatial object location in our system. In our system spatial object location in an image is used as the global feature. The object's mutual spatial relationship is calculated based on the algorithm adopted from the concept of principal component analysis (PCA), proposed by Chang and Wu [14] and later modified by Guru and Punitha [15], to determine the first principal component vectors (PCVs).
3. to help the user ask a query in GUI. The user chooses for a query graphical objects semantically collected in groups.
4. compare image objects coming from the same class as a stage in the image retrieval process. Let a query be an image I_q, such as $I_q = \{o_{q1}, o_{q2},\ldots, o_{qn}\}$. An image in the database will be denoted as I_b, $I_b = \{o_{b1}, o_{b2},\ldots, o_{bm}\}$. In order to answer the query, represented by I_q, we compare it with each image I_b in DB. We determine the similarity between vectors of their signatures. Next, we find the spatial similarity between their PCVs. Later, we proceed to the final step, namely, we compare the similarity of the objects representing both images I_q and I_b, respectively between objects of the same class.

1.6 Conclusions

In this chapter, first we have determined the ability of fuzzy sets and fuzzy rule-based classifiers to classify graphical objects in our CBIR system. We have shown an example of classification based on nine and three fuzzy rules according to the data character. We have chosen the most distinguished coordinates from a feature vector in order to exemplify the proposed method that seems to be quite promising.

Intensive computational experiments are under way in order to draw some conclusions regarding the choice of parameters for the model, including the choice of the above-mentioned metrices. However, the preliminary results we have obtained so far, using the simplest configuration, are quite hopeful.

As for the prospects for future work, the implementation of an optimised procedure should verify the feasibility of the approach. We expect a reasonable performance from the evaluation strategy outlined in the chapter.

References

[1] Deb, S. (ed.): Multimedia Systems and Content-Based Image Retrieval, ch. VII and XI. IDEA Group Publishing, Melbourne (2004)

[2] Ali, J.M.: Content-Based Image Classification and Retrieval: A Rule-Based System Using Rough Sets Framework. In: Ma, Z. (ed.) Artificial Intelligence for Maximizing Content Based Image Retrieval, New York, ch. IV, pp. 68–82 (2009)

[3] Niblack, W., Flickner, M., et al.: The QBIC Project: Querying Images by Content Using Colour, Texture and Shape. In: SPIE 1908, pp. 173–187 (1993)

[4] Ogle, V., Stonebraker, M.: CHABOT: Retrieval from a Relational Database of Images. IEEE Computer 28(9), 40–48 (1995)

[5] Pons, O., Vila, M.A., Kacprzyk, J.: Knowledge management in fuzzy databases. STUDFUZZ, vol. 39. Physica–Verlag, New York (2000)

[6] Lee, J., Kuo, J.-Y., Xue, N.-L.: A note on current approaches to extending fuzzy logic to object oriented modeling. International Journal of Intelligent Systems 16(7), 807–820 (2001)

[7] Berzal, F., Cubero, J.C., Kacprzyk, J., Marin, N., Vila, M.A., Zadrożny, S.: A General Framework for Computing with Words in Object-Oriented Programming. In: Bouchon-Meunier, B. (ed.) International Journal of Uncertainty. Fuzziness and Knowledge-Based Systems, vol. 15(suppl.), pp. 111–131. World Scientific Publishing Company, Singapore (2007)

[8] Ma, Z.M., Zhang, W.J., Ma, W.Y.: Extending object-oriented databases for fuzzy information modeling. Information Systems 29, 421–435 (2004)

[9] Cubero, J.C., Marin, N., Medina, J.M., Pons, O., Vila, M.A.: Fuzzy Object Management in an Object-Relational Framework. In: Proceedings of the 10th International Conference IPMU, Perugia, Italy, pp. 1775–1782 (2004)

[10] Candan, K.S., Li, W.-S.: On Similarity Measures for Multimedia Database Applications. Knowledge and Information Systems (3), 30–51 (2001)

[11] Jaworska, T.: Object extraction as a basic process for content-based image retrieval (CBIR) system. Opto-Electronics Review, Association of Polish Electrical Engineers (SEP) 15(4), 184–195 (2007)

[12] Jaworska, T.: Database as a Crucial Element for CBIR Systems. In: Proceedings of the 2nd International Symposium on Test Automation and Instrumentation, vol. 4, pp. 1983–1986. World Publishing Corporation, Beijing (2008)

[13] Chang, C.C.: Spatial match retrieval of symbolic pictures. J. Informat. Sci. Eng. 7, 405–422 (1991)

[14] Chang, C.C., Wu, T.C.: An exact match retrieval scheme based upon principal component analysis. Pattern Recognition Letters 16, 465–470 (1995)

[15] Guru, D.S., Punitha, P.: An invariant scheme for exact match retrieval of symbolic images based upon principal component analysis. Pattern Recogn. Lett. 25, 73–86 (2004)

[16] Zadeh, L.A.: Fuzzy sets. Information and Control 8(3), 338–353 (1965)

[17] Ishibuchi, H., Yamamoto, T.: Rule weight specification in fuzzy rule-based classification systems. IEEE Transactions on Fuzzy Systems 13(4), 428–435 (2005)

[18] Mozaki, K., Ishibuchi, H., Tanaka, H.: Adaptive fuzzy rule-based classification systems. IEEE Transactions on Fuzzy Systems 13(4), 238–250 (1996)

[19] Ishibuchi, H., Nojima, Y.: Toward Quantitative Definition of Explanation Ability of fuzzy rule-based classifiers. In: IEEE International Conference on Fuzzy Systems, Taipai, Taiwan, June 27-39, pp. 549–556 (2011)

[20] Teague, M.R.: Image analysis via the general theory of moments. In: JOSA, 8th edn., vol. 70, pp. 920–930 (1980)

[21] Jaworska, T.: A Search-Engine Concept Based on Multi-feature Vectors and Spatial Relationship. In: Christiansen, H., De Tré, G., Yazici, A., Zadrozny, S., Andreasen, T., Larsen, H.L. (eds.) FQAS 2011. LNCS(LNAI), vol. 7022, pp. 137–148. Springer, Heidelberg (2011)

Chapter 2
Decentralized Multi-Agent Algorithm for Translational 2D Image Alignment

Tomáš Machálek and Kamila Olševičová

Abstract. We present a novel multi-agent algorithm applied to the problem of image alignment. Our method operates with multiple concurrent solutions held by agents who each attempt to reach the lowest error function score by trying to place a segment from a translated image to an unsegmented fixed image. Agents borrow and return segments of the translated image from a shared repository and iteratively suggest and evaluate their particular solutions. Finally, the global solution is determined by clustering of agents' individual results. Experiments show that our approach provides results of high reliability and performance compared with traditional intensity based registration methods that rely on global optimization of a single error function given by translation of whole image.

2.1 Introduction

Image alignment in general is a task of transforming two or more images in such a way that they can be placed into a common coordinate system with respect to the scene they depict. Until now many approaches have been proposed and used in areas such as medical imaging, video stabilization, remote sensing and many others [1, 2]. So called intensity (or area) based methods use specific measure (e.g. sum of absolute differences, cross correlation, mutual information) to quantify a similarity of overlapping regions using all participating pixel intensity values. Then some optimization technique is involved to find a transformation giving the highest similarity.

Selecting proper optimization method depends significantly on the nature of problem. To correct some minor error in mutual position of two images, we can

Tomáš Machálek · Kamila Olševičová
University of Hradec Králové, Faculty of Informatics and Management,
Dept.of Information Technologies, Hradec Králové, Czech Republic
e-mail: {Tomas.Machalek,Kamila.Olsevicova}@uhk.cz

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 15–24.
springerlink.com

apply local optimization or even exhaustive search, because the search space is small compared with the size of search space given by the issue of searching across all translations resulting in non-empty intersection of the images. But except for such small search spaces, the shape of the objective function is typically complex, containing many local optima, which places high demands on optimization method.

In general, global optimization problems solved on discrete, bounded search spaces are guaranteed to be solved only by a full-space search, which is usually unacceptable due to its time and/or memory complexity. In practice this fact makes us to look for methods with limited accuracy and reliability but with more realistic computational complexity.

Many global or pseudo-global optimization methods have been used along with image registration - Simulated Annealing [3], Nelder-Mead simplex method [4], Powell's method [5] or Particle Swarm Optimization [6, 7] to name a few. In our previous experiments [7] we found Nelder-Mead method combined with random restarts and Particle Swarm Optimization giving promising results of about same quality and performance. But even these algorithms when facing large search spaces and complex images (like satellite images or images containing repeating structures) tend to be trapped in a local optimum.

There are several approaches to deal with this problem in case of intensity based methods. A common solution is to process images represented in multiple scales [1], [2]. Typical procedure starts with the lowest resolution where the level of detail is reduced, resulting in one or more possible solutions which can be confirmed or disproved when moving to higher resolution levels where only local search is performed. This general "processing template" can be combined with numerous image registration algorithms, but again, in general it does not guarantee error-free results [1, 2].

Another possible way to overcome mentioned issues lies in performing multiple "sub-registrations" using segments of original images and performing a synthesis of these particular tasks' results. For example in [8] and [9] authors developed a multi-agent systems with certain common characteristics. Both fixed and transformed images are split into smaller segments, multiple types of agents with specific tasks are present and some form of coordination, or even central control, exists - former solution uses blackboard architecture while the latter is controlled by a supervising agent.

Unlike in case of a single result methods, multiple, simultaneous results lead also to multiple possible interpretations, where selecting best score represents only one of possible ways how to infer a single, hopefully the best possible value.

The structure of the chapter is as follows. In section 2.2 the problem is specified. In section 2.3 principles of the proposed algorithm are explained. Section 2.4 deals with complexity estimation. Experiments are discussed in section 2.5 and further research directions are suggested in section 2.6.

2.2 Problem Statement

Considering two raster images I_1 (w_1 x h_1 pixels), I2 (w_2 x h_2 pixels) placed in 2D Euclidean space where I_2 can be moved along both axes while I_1 is fixed, we can reformulate the problem as a parameterized relationship between two functions from R^2. We assume that any value outside of areas of processed images has some constant intensity value - for our text it is 0. Then we define general error function on some search space $A: (u, v) \in [u_1, u_2] \times [v_1, v_2]$ where $u_1, u_2, v_1, v_2 \in \mathbb{R}$:

$$E(u, v) = \iint_{-\infty}^{\infty} f(I_1(x, y), I_2(x + u, y + v))dxdy \tag{2.1}$$

It is reasonable to restrict the integration to a finite area where at least one image function is non zero. In practice the search space *A* must be selected carefully, because getting too small overlapping regions on original finite images may lead to biased results and thus incorrect alignment [2]. This can be avoided either by selecting smaller search space or by penalizing the cost function value in some regions [2].

According to the previous, the alignment problem can be formulated as

$$\left(u_{opt}, v_{opt}\right) = argmin_{(u,v)}E(u, v) \tag{2.2}$$

For our experiments we have been using *sum of absolute differences* and *normalized cross correlation* as an *E* function.

2.3 Proposed Solution

Note: when speaking about a *random* selection or position, we have the uniform distribution in mind.

Instead of trying to position properly image I_2 as a whole, our algorithm operates with many smaller segments generated from I_2 and searches best location for each of them according to the selected function *E*. Image I_1 is not segmented.

Initial configuration can be seen on fig. 2.1. All *n* segments are of the same size and are obtained from randomly selected locations of I_2. In fact, we have been using two strategies for this step:

1. *n* randomly selected rectangles within I_2,
2. $\lceil n * r \rceil$, ($r > 1$) randomly picked candidate rectangles; *n* of them with the highest entropy selected for further processing; we refer to this method as to the "entropy hint" in further text.

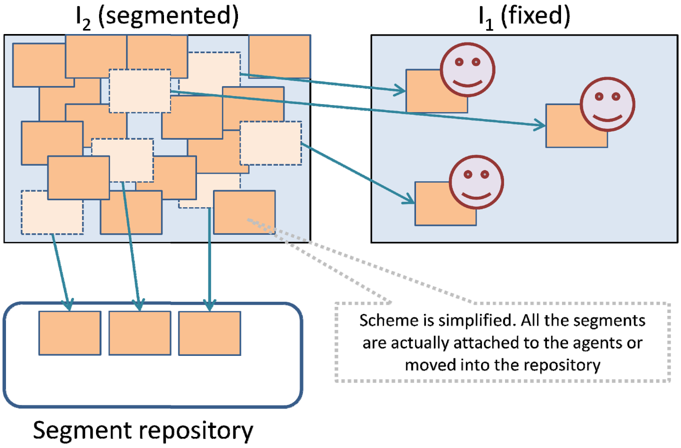

I2 (segmented)
I1 (fixed)
Scheme is simplified. All the segments are actually attached to the agents or moved into the repository
Segment repository

Fig. 2.1 Simplified scheme of the initialization. Random segments are generated from I_2, n_a of them is attached to the agents and the rest is stored in the segment repository.

Our selection process in general does not guarantee that the set of n segments will cover all the information that I_2 provides and on the other side at least some of the segments may overlap each other and thus carry some level of redundancy. Level of this phenomenon depends on size of I_2, number of the segments and their size. For example in case of 640x480 pixel image, when placing 50 segments of size 40x40 pixels (which was one of our typical configurations), the coverage was approximately 40% while 20% of the area was covered by more than one segment.

In the next step, n_a agents are created. Their coordinate system is defined according to the area of the image I_1 and their initial position is set randomly. Each agent selects one random segment which along with agent's position becomes its initial solution. The value of n_a must comply the equation $n = n_a * (s_e + 1)$. The constant $s_e \in \mathbb{N}$ is explained later.

Agent's calculation is performed iteratively with predefined number of iterations. Each iteration consists of two phases:

1. exploration (fig. 2.2),
2. evaluation (fig. 2.3).

During the exploration phase agent borrows random se segments from the segment repository and along with his own segment he generates $s_e + 1$ random coordinates (x_i, y_i) in his neighborhood. Size of this neighborhood is decreasing progressively with each iteration i according to the function

$$g(i) = \left[\frac{w_2}{a + \delta \cdot i}, \frac{h_2}{b + \delta \cdot i}\right], a, b > 0, \delta > 0, i \in \mathbb{N}_0 \qquad (2.3)$$

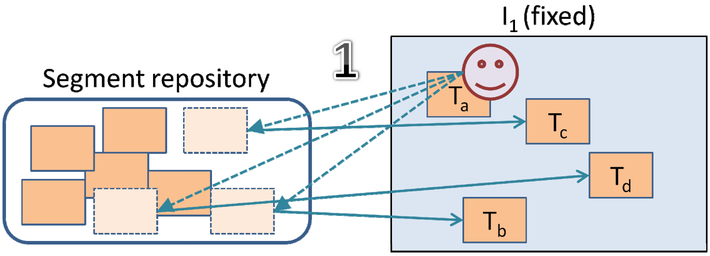


Fig. 2.2 Single agent iteration – the exploration phase. (1) Agent selects a small random group of unused segments from the repository and places them randomly with respect to the current range *g* (see equation 2.3) which decreases gradually.

The parameters *a*, *b* and *δ* of the function (3) must be selected carefully because too small initial range may prevent algorithm from searching all requested transformation space. In other words, we must assume a worst case scenario when all agents start to search being positioned in the outermost locations according to the destination locations of the segments they carry.

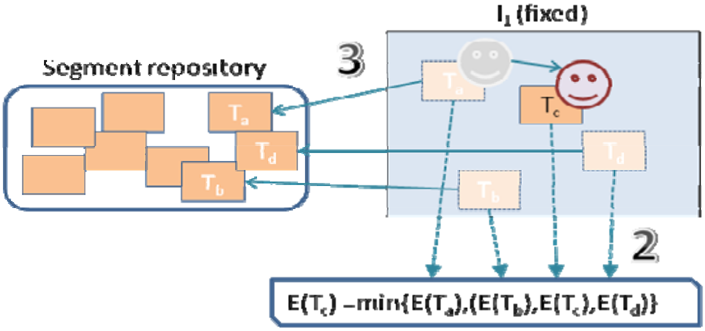


Fig. 2.3 Single agent iteration - the evaluation phase. (2) Agent chooses best segment as the one he carries (here T_c) and moves to its position, (3) the rest of the segments is returned to the repository.

Depending on his position and the current value of the range g, agent may generate coordinates reaching (partially or completely) out of the image I_1. In such case the value must be corrected to keep tested segment within the I_1. This can be accomplished e.g. by decreasing of the absolute value of affected coordinate or by calculating the coordinate using modulo arithmetic.

In the evaluation phase the agent compares the scores reached by $s_e + 1$ segments he placed randomly, where the scores are results of used error function (e.g. in our case the cross correlation or sum of absolute differences). Then the agent selects the best segment as the one he carries and moves to this segment's position. Remaining segments are returned to the repository and are immediately available to be borrowed by any agent in the next iteration. Both described phases can be seen in form of a Python code on fig. 2.4.

After the defined number of iterations, final evaluation is necessary as we have n_a agents, each holding his own solution. To have a robust method, we generally cannot expect that all (or at least significant number of) agents will agree on a single solution or that the best agent will always represent correct solution.

On the contrary, we have to count with the fact that many of the results, even some results with top scores will be wrong when applied to the original task. Based on empirical results we have proposed following heuristics:

1. we should be able to detect the solution using the best 10% of agents,
2. global best solution is the one that most of these 10% agents agreed on (with some defined tolerance).

We have implemented this part as a clustering task where the number of clusters is not known. A hierarchical clustering function from the SciPy library *(scipy.cluster.hierarchy.fclusterdata)* was used for this purpose.

```
def update_agent(agent, repository, n_seg, i):
    """
    agent: object representing an agent
    repository: unused segments repository
    n_seg: number of segments to be tested along with
agent's
    own segment
    i: iteration number
    """
    tested_segments = [agent.segment]
+ fetch_random_segments(n_seg, repository)
    best_val = None
    best_pos = None
    best_segment = None
    for segment in tested_segments:
        new_pos = generate_random_coords(i, agent.pos)
        val = evaluate_segment(new_pos, segment)
        if val < best_val or best_val is None:
            best_val = val
            best_segment = segment
            best_pos = new_pos
    agent.segment = best_segment
    agent.pos = best_pos
    for segment in tested_segments: # return other
segments
        if segment is not best_segment:
            repository.append(segment)
```

Fig. 2.4 Exploration and evaluation phases as a snippet of Python code

2.4 Estimation of Computational Complexity

For simplicity, let us assume that we register smaller image I_2 (size $w_2 \times h_2$) against bigger image I_1 and that for all explored transformations, the area of I_2 lies completely within the area of I_1.

Having na agents, each evaluating se segments of size $s_w \times s_h$ pixels leads to the following calculation expressing the total number of pixel evaluations needed to process a single iteration:

$$N_{pcalc} = n_a \cdot s_e \cdot s_w \cdot s_h \tag{2.4}$$

Using traditional approach, here with the Particle Swarm Optimization as a global optimizer and assuming np particles, for a single iteration we get:

$$N_{pcalc} = n_p \cdot w_2 \cdot h_2 \tag{2.5}$$

Considering our typical testing set-up where $n_a = 50$, $s_e = 4 + 1$ for our algorithm and $n_p = 30$ for the PSO algorithm and assuming the same number of iterations for both methods, we get (making right-hand sides of 2.5 and 2.6 equal) maximum segment size for our method to be able to compete with the PSO as

$$s_w \cdot s_h = 0.12 \cdot w_2 \cdot h_2 \tag{2.6}$$

which means that for an image of size 640x480 pixels, square segments must be of size 192x192 pixels or less to be comparable with the PSO set-up in terms of calculation time. Since our typical segment size was 60x60 pixels for such image, then according to the equation (2.6), the number of pixel calculations for *one iteration* was approximately ten times smaller.

In practice, the calculation times were rather of similar length. So far we have been able to identify two possible causes. First, some translations PSO searched through led to only partial images overlapping and thus the number of pixel calculations was smaller in these cases. Second, the PSO variant calculates single "large" overlapping at once, which leads to optimized, vectorized calculation using *SciPy* and *NumPy* functions, while our solution distributes calculation in $n_a \cdot s_e$ smaller chunks which adds some overhead when calculated in a non-parallel manner.

2.5 Experiments

Testing environment was realized using Python language along with packages *NumPy*, *SciPy* and *mahotas*. We have been testing numerous images in two main scenarios:

1. two images copied from two mutually shifted rectangular areas of a single original image; registration performed using the *sum of absolute differences* error function,
2. the same approach as in previous case but I_2 was modified by the *posterize* (level of posterization 4) effect using the Gimp image editor; registration performed using *normalized cross-correlation* error function.

We have been measuring quality of the solution represented by the best result's distance to the real optimum. The standard deviation of this value in repeated experiments indicated how stable the configuration was. We have been also monitoring time necessary to calculate the solution. The results we have achieved after repeated experiments can be seen on figures 2.5, 2.6, 2.7 and 2.8.

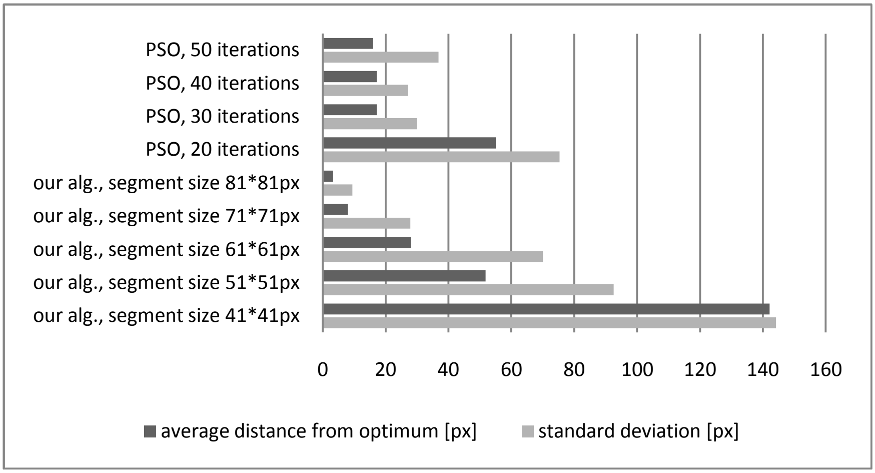


Fig. 2.4 Traditional area-based registration of two satellite images using normalized cross correlation along with the Particle Swarm Optimization algorithm (30 particles), compared with our algorithm (50 iterations, 50 agents, 4 new segments for an agent). The values are averages from 50 measurements.

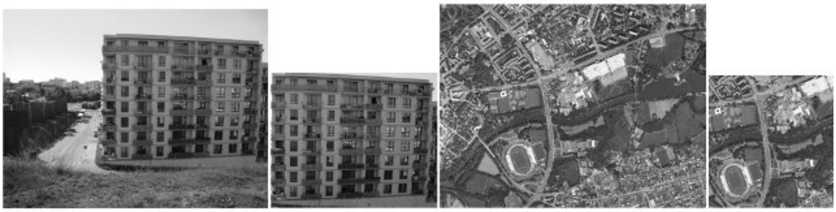


Fig. 2.5 Images used as a source of presented results. The leftmost image pair is referred as "A", the rightmost one as "B"

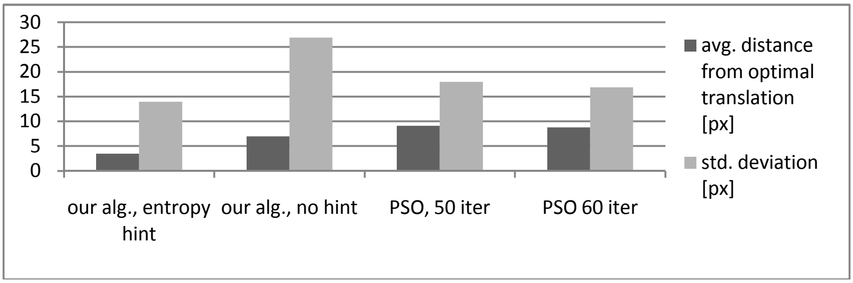


Fig. 2.6 Solution quality compared. SAD similarity metrics, image pair "A", 40 agents, 50 iterations, segment size 37x37 pixels, both entropy hint and no hint used for segment selection in case of our algorithm; for PSO algorithm there were 30 particles and 50 and 60 iterations used. The values are averages from 50 measurements.

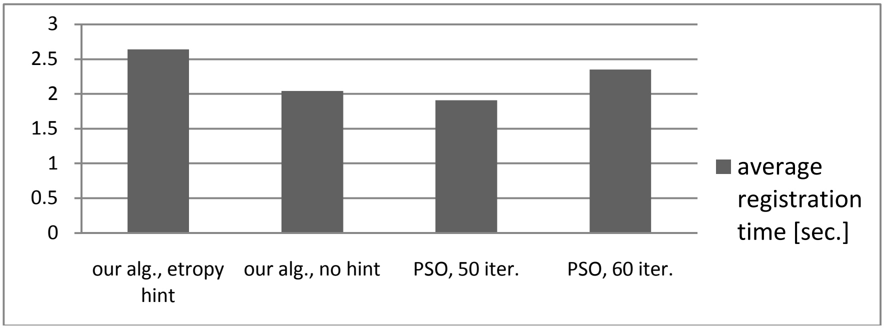


Fig. 2.7 Calculation times compared. SAD similarity metrics, image pair "A", 40 agents, 50 iterations, segment size 37x37 pixels, both entropy hint and no hint used for segment selection in case of our algorithm; for PSO algorithm there were 30 particles and 50 and 60 iterations used. The values are averages from 50 measurements.

2.6 Conclusions

Image alignment using multi-agent system was presented. The experiments showed that we had been able to obtain results of higher reliability compared with Particle Swarm Optimization based solution. A prototype of the algorithm is available as a Python project on address http://code.google.com/p/marg/ to allow further development and reproducing of presented results.

For further research we would suggest especially examination of the effect the algorithm's parameters have on the calculation progress and result. Also the ability of the method to deal with some real use scenarios, where a wide range of error functions and images is used, should be studied.

Acknowledgement. The support of the Czech Science Foundation under Grant 402/09/0405 is kindly acknowledged.

References

1. Zittová, B., Flusser, J., Šroubek, F.: Image registration: A survey and recent advances (2005)
2. Szeliski, R.: Computer Vision: Algorithms and Applications. Texts in Computer Science. Springer (2010)
3. Ritter, N., Owens, R., Cooper, J., et al.: Registration of stereo and temporal images of the retina. IEEE Transactions on Medical Imaging 18, 404–418 (1999)
4. Holia, M., Thakar, V.K.: Image registration for recovering affine transformation using nelder mead simplex method for optimization. Computer Science Journals 2009 (2009)
5. Čapek, M.: Optimisation strategies applied to global similarity based image registration methods (1999)
6. Das, A., Bhattacharya, M.: Affine-based registration of ct and mr modality images of human brain using multiresolution approaches: comparative study on genetic algorithm and particle swarm optimization. Neural Computing & Applications 20, 223–237 (2011)
7. Machálek, T.: Application of particle swarm optimization in 2d image alignment, Master's thesis, University of Hradec Králové (in Czech) (2011)
8. Tait, R.J., Schaefer, G., Hopgood, A.A.: Intensity-based image registration using multiple distributed agents. Know.-Based Syst. 21, 256–264 (2008)
9. Amine Jallouli, M., Zagrouba, E., et al.: Decomposition of an alignment problem of two 3d images by a multi-agent approach. Innovations in Information Technology, 680–684 (2007)

Chapter 3
A Sparse Reconstruction Approach to Video Deinterlacing

Maria Trocan

Abstract. With the apparition of digital television and flat displays, interlaced to progressive frame format conversion represents an importantant video systems feature. In this chapter, we use an inverse problem formulation for video deinterlacing and propose a two-step sparse-reconstruction algorithm for solving it. Firstly, an edge-preserving approximation of the progressive frame is obtained and used for triggering a bidirectional motion-compensated prediction for the current field. In a second step, a sparse residual is calculated as difference between the current field and the projection of its temporal prediction using the same parity sampling matrix. This field residual is further reconstructed using a total-variation regularization method and added back to the motion-compensated prediction to form the final progressive frame. The proposed deinterlacing method presents high quality results compared to other deinterlacing approaches.

3.1 Introduction

Introduced by the old analog television transmission systems as a trade-off between framerate and bandwidth capacity, the interlaced video format has become obsolete today, when all transmissions are digital. Nowadays, almost all displays - whether LCD, plasma or LED, as well as the video encoders, require progressive video input, whereas much of the available video content is in interlaced format. In order to solve this problem several deinterlacers, varying in quality and required computational power, have been proposed in the last years.

Usually, the lowest complexity is attached to spatial deinterlacers, which use only the information in the current field to interpolate the missing one [1]. However, these

Maria Trocan
Institut Supérieur d'Électronique de Paris,
Signal and Image Processing Department,
28 rue Notre Dame des Champs, Paris, France
e-mail: maria.trocan@isep.fr

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 25–33.
springerlink.com

deinterlacing methods are not optimal, as neither temporal information, nor motion activity is considered in the interpolation. Moreover, they fail to remove the flicker artifacts within dynamic areas.

To alleviate this issue, motion-compensated (MC) deinterlacers methods have been proposed [2, 3], interpolating thus the missing field along the estimated motion trajectory. Generally speaking, MC-algorithms outperform the other approaches at a higher computational cost. In order to lower this complexity, most MC implementations [4, 5] use block-based motion-estimation (ME). However, these approaches introduce blocking artifacts, which are more visible on highly textured areas. Further, unreliable motion information may limit the use of MC techniques. For these reasons, different hybrid methods have been proposed [6, 7], combining directional interpolation and motion compensation techniques.

Lately, total-variation (TV)-based reconstruction techniques have been proposed for video deinterlacing. In [8, 9], Keller *et al.* propose to use the inpainting TV-reconstruction solution in an adaptive MC-deinterlacer and in [10], TV-regularization with spatio-temporal smoothness constraints is employed for interlaced to progressive frame conversion. The TV-based interpolators present smooth image reconstructions, but however, the high-complexity attached these algorithms prohibit their use in real-time deinterlacing setups.

In this chapter we propose a deinterlacing scheme which takes advantage of the TV-based reconstruction in a motion-compensated context, and enhances the frame recovery accuracy by applying the TV-reconstruction on a sparse field representation. In the sequel, deinterlacing is formulated as an inverse reconstruction problem and, rather than spatially reconstructing the missing fields, the proposed method takes advantage of the temporal correlation on the motion direction. A sparse, field-parity-coherent residual is obtained as difference between the current field and its bidirectional motion-compensated predictor, and a smooth-gradient TV-regularization is used for the reconstruction of this residual. The final progressive frame is given by adding the resulted progressive format residual to the original temporal prediction.

The proposed algorithm results in high quality, smooth progressive frame conversion, by alleviating the motion-artifacts due to the gradient penalty TV-reconstruction of the MC-residual. Moreover, in our framework we consider a reduced-complexity implementation of TV-regularization [11], lowering thus the computational burden attached to this image reconstruction method.

This chapter is organized as follows: in Section 3.2 we introduce the sparse reconstruction problem and the TV-regularization solution, before presenting the proposed deinterlacing method in Section 3.3. The experimental results, presenting the performance of our approach in comparison to other deinterlacing methods, are given in Section 3.4. Finally, conclusions are drawn in Section 3.5.

3.2 Background

Over the last years, sparse reconstruction schemes have become very popular for the solution of linear and nonlinear inverse problems [12], such as image inpainting, super-resolution or compressed sensed acquisition. In these inverse problems, a real-valued signal x of length N has to be recovered from a subset of M samples. In other words, x should be reconstructed from a partial observation $y = \boldsymbol{\Phi} x$, where y has length M, and $\boldsymbol{\Phi}_{M \times N}$ is called the sub-sampling or measurement matrix.

As $\hat{x} = \boldsymbol{\Phi}^{-1} y$ reconstruction is ill-posed, signal recovery is possible if x is sufficiently sparse in a certain space. In this case, the sparsity condition for x recovery will exist with respect to some unknown transform $\boldsymbol{\Psi}$ and the reconstruction process is resumed to the production of a sparse set of significant transform coefficients, $\hat{x} = \boldsymbol{\Psi} x$. The recovery procedure searches for $\hat{x}$ with the smallest l_0 norm consistent with the observed y:

$$\hat{x} = \arg\min_{\hat{x}} \|\hat{x}\|_0 , \quad \text{such that} \;\; y = \boldsymbol{\Phi}\boldsymbol{\Psi}^{-1}\hat{x} \tag{3.1}$$

where $\boldsymbol{\Psi}^{-1}$ represents the inverse transform. Due to NP-completeness of this l_0 optimization, alternative procedures have been proposed [12] for sparse reconstructions using l_1 or l_2-norms .

Total variation (TV) minimization [13] have been extensively used in image reconstruction problems [14, 15, 11]. TV-based reconstruction methods replace the search of the sparsest solution within the transform $\boldsymbol{\Psi}$ domain with the smoothest solution within the space of possible solutions. This is possible by using the ℓ_1 norm to enforce sparsity upon the gradient of the searched solution, creating thus a penalty function of the form:

$$TV(x) = \sum_i \sum_j \left| x_{i+1,j} - x_{i,j} \right| + \left| x_{i,j+1} - x_{i,j} \right| . \tag{3.2}$$

Using the above cost function, the image-recovery problem can be stated as:

$$\hat{x} = \arg\min_{x} \; ||y - \boldsymbol{\Phi} x||_2 + \lambda \; TV(x). \tag{3.3}$$

TV minimization has been widely used in sparse reconstructions; some computationally efficient approaches to solving (3.3) have been proposed, such as iterative soft thresholding [14], alternating minimization [15] or the alternating direction algorithm (TVAL3) proposed in [11].

Among these, TVAL3 method has a decreased computational burden, compared to other TV minimization approaches for image reconstruction. Moreover, it provides high-quality reconstructed images, therefore we propose to use it in our sparse-reconstruction deinterlacing framework.

In the sequel we explain how sparse-reconstruction schemes can be used for interlaced to progressive conversion and introduce the proposed motion-compensated residual frame-recovery approach.

3.3 Proposed Method

Deinterlacing can be translated into an inverse reconstruction problem, wherein the real-valued signal x represents the progressive $N_1 \times N_2 = N$ frame, and the sub-sampling matrices used for obtaining the odd, $\Phi^{t=2k+1}$, and even, $\Phi^{t=2k}$, interlaced fields at alternating parity time instants t are given by:

$$\Phi_{i,j}^{t=2k+1} = \begin{cases} 1, & if \quad j = 2i-1 \\ 0, & otherwise, \end{cases} \tag{3.4}$$

$$\Phi_{i,j}^{t=2k} = \begin{cases} 1, & if \quad j = 2i \\ 0, & otherwise, \end{cases} \tag{3.5}$$

for $\forall i = 1 \dots N_1/2, \quad j = 1 \dots N_2$.

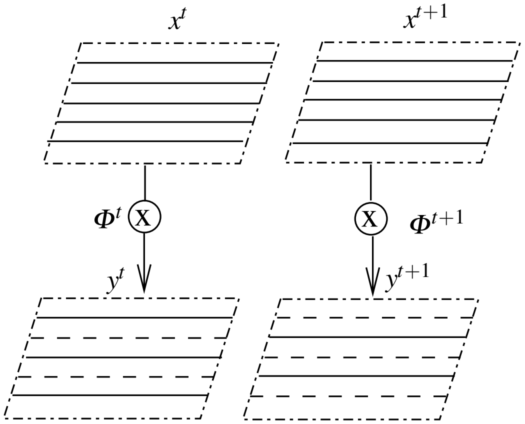


Fig. 3.1 Progressive to interlaced frame format conversion.

The interlaced field y^t is therefore the result of sub-sampling the progressive frame x^t with the corresponding parity matrix Φ^t, i.e.: $y^t = \Phi^t x^t$, as described in Fig. 3.1.

In the followings we propose an algorithm which incorporates motion estimation and compensation into the TV-based recovery [11] with the goal of improving the deinterlacing quality on the temporal direction. In the sequel, we refer to this proposed approach as sparse reconstruction deinterlacing (SRD).

The SRD algorithm, described in Fig. 3.2, is partitioned into two steps. In the first step, an edge-preserving recovery of the current field $\mathbf{y}^t$ and its temporal neighbours, $\mathbf{y}^{t-1}$ and $\mathbf{y}^{t+1}$, is obtained by spatially interpolating them using the edge line averaging algorithm (ELA):

$$\tilde{x}_{i,j}^t = \frac{y_{i-1,j+x_0}^t + y_{i+1,j-x_0}^t}{2}. \tag{3.6}$$

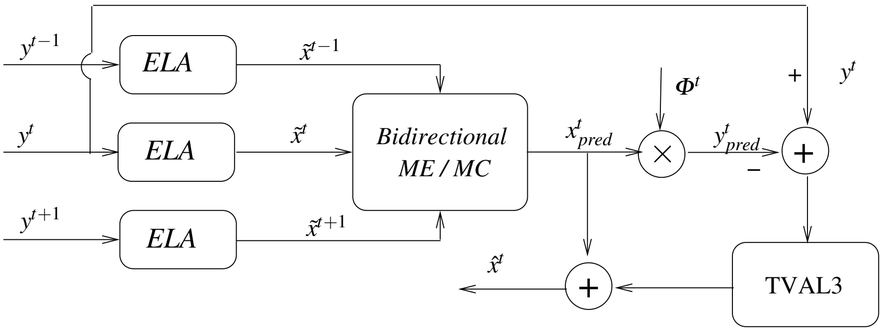


Fig. 3.2 Block-scheme of the proposed SRD deinterlacing algorithm.

In (3.6), the exact value of x_0 is given by performing a minimization on three edge directions:

$$\left|y^t_{i-1,j+x_0} - y^t_{i+1,j-x_0}\right| = \min_{x_0\in\{-1,0,1\}} \left|y^t_{i-1,j+x_0} - y^t_{i+1,j-x_0}\right|. \tag{3.7}$$

Note that in this first step any other spatial interpolation method can be used for obtaining a first reconstruction $\tilde{x}$ of the progressive frame. In our framework we employed ELA due to its low-complexity implementation.

Following a bidirectional ME/MC between the current frame reconstruction $\tilde{x}^t$ and the left, $\tilde{x}^{t-1}$, and right, $\tilde{x}^{t+1}$, neighbors, a motion-compensated prediction of the current frame x^t_{pred} is created as:

$$\begin{aligned} x^t_{pred}(i,j) = 0.5\tilde{\mathbf{x}}^{t-1}(i - MVx^t_{left}, j - MVy^t_{left}) + \\ 0.5\tilde{\mathbf{x}}^{t+1}(i - MVx^t_{right}, j - MVy^t_{right}), \end{aligned} \tag{3.8}$$

where MVx^t / MVy^t represents the vertical/horizontal components of the backward MV_{left} / forward MV_{right} motion vectors for the current reconstruction of the frame $\tilde{x}^t$.

In a second step, we compute a sparse MC-residual r^t for the current field as difference between y^t and the same parity sub-sampled predictor $y^t_{pred} = \Phi^t x^t_{pred}$. This residual:

$$r^t = y^t - y^t_{pred}, \tag{3.9}$$

is then reconstructed using the TVAL3 method in [11] and added back to x^t_{pred} to obtain the final deinterlaced image $\hat{x}^t$, i.e.:

$$\hat{x}^t = x^t_{pred} + TVAL3(r^t). \tag{3.10}$$

It should be noted that a spatial TV-reconstruction method [14, 15, 11] can be directly used for solving (3.3) and therefore providing a direct progressive frame recovery $\hat{x}^t$. A comparison of our proposed method with the direct TV-reconstruction in [11] is presented in our simulation framework. As the quality of signal recovery is highly related to the sparsity of the signal to be reconstructed [16], our deinterlacing method uses the recovery of a much sparser signal, e.g. the motion compensated residual r^t, rather than the field y^t. Moreover, due to the smoothing properties of TV-based reconstruction, it is very important to preserve the edge content and therefore guarantee a high quality temporal prediction x^t_{pred}, hence the use of ELA as spatial interpolator and the high accuracy method proposed in [17] for motion estimation.

3.4 Experimental Results

In our experimental framework, we consider several CIF-352 × 288 (”Foreman”, ”Hall”, ”Mobile” and ”Stefan”) and one QCIF-176 × 144 (”Carphone”) video sequences for testing the proposed method. These sequences, which have been chosen for their different texture content and motion dynamics, were originally in progressive format.

The interlaced content have been obtained by multiplying the original frames with the odd/even sampling matrix $\boldsymbol{\Phi}$, and thus the even lines of the even frames and the odd lines of the odd frames were removed as shown in Fig. 3.1. This way, objective quality measurements could be done, using the original sequences - progressive frames - as references. The tests were run on 50 frames for each sequence.

The deinterlacing performance of our method is presented in terms of peak signal to-noise ratio (PSNR, in Table 3.1) computed on the luminance component. The proposed algorithm (SRD) is compared to Vertical Average (VA), Edge Line Average (ELA), Temporal Field Average (TFA), Adaptive Motion Estimation (AME) and Motion-Compensated Deinterlacing (MCD), which are the most common

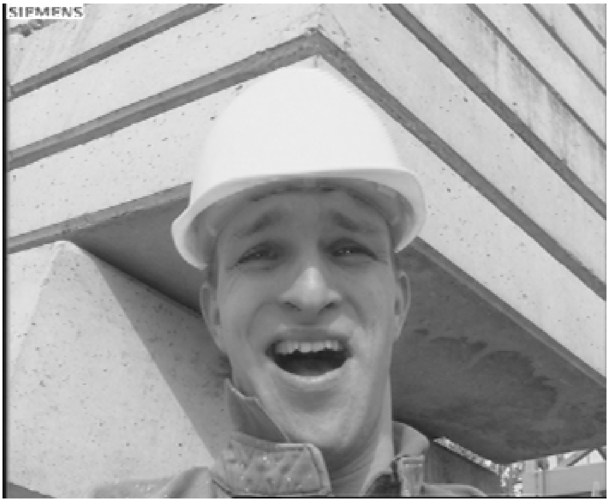


Fig. 3.3 SRD reconstruction of the 26st frame from ”Foreman” CIF sequence.

Table 3.1 PSNR (dBs) comparison of the proposed SRD algorithm with classical deinterlacing methods.

	Foreman	Hall	Mobile	Stefan	Carphone
VA	32.15	28.26	25.38	27.30	32.17
ELA	33.14	30.74	23.47	26.04	32.33
TVAL3 [11]	29.37	26.87	23.17	24.25	29.84
TFA	34.08	37.47	27.96	26.83	37.39
AME	33.19	27.27	20.95	23.84	29.63
MCD	35.42	34.23	25.26	27.32	33.55
EPMC	37.18	39.08	30.56	30.11	37.55
SMCD	37.52	**39.71**	30.41	31.77	37.59
SRD	**37.64**	39.66	**31.62**	**32.57**	**39.42**

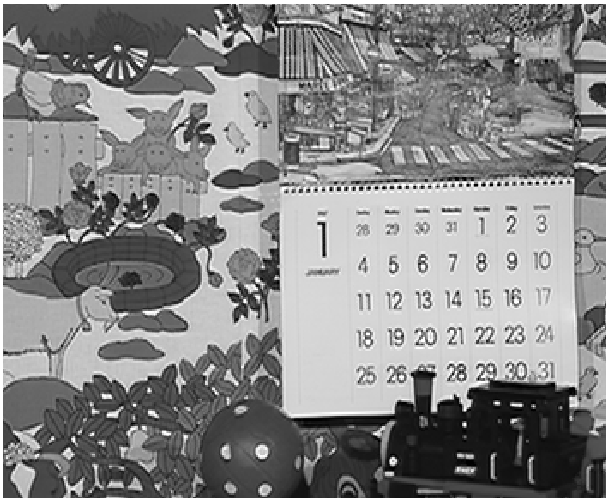


Fig. 3.4 SRD reconstruction of the 49st frame from ”Mobile” CIF sequence.

implementations in deinterlacing systems. Moreover, the efficiency of SRD is compared to our previous methods in [6, 7], denoted by EPMC, respectively SMCD.

SRD has the best performance for all tested sequences and visually it results in smooth-deinterlaced content, outperforming the classical deinterlacing methods with ≈ 5 dBs in average. Moreover, the proposed approach has an average PSNR gain of ≈ 0.5 dBs (2 dBs for ”Carphone”) with respect to the deinterlacing methods in [6, 7] and a significant gain of ≈ 9 dBs with respect to the direct TV-based reconstruction in [11].

A visual evaluation of deinterlacing outcome for SRD is proposed in Figures 3.3 - 3.5, for three of the considered test sequences.

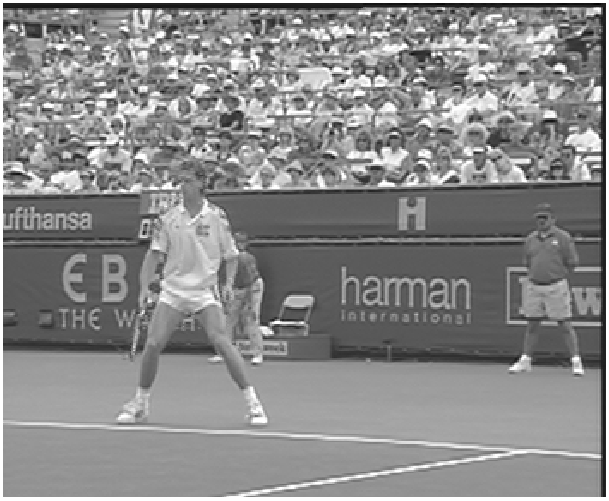


Fig. 3.5 SRD reconstruction of the 45^{st} frame from ”Stefan” CIF sequence.

3.5 Conclusion

In this chapter, deinterlacing is formulated as an inverse optimization problem and a sparse-reconstruction solution is proposed for solving it. In order to take advantage of the inter-field correlation and to enforce the sparsity of the signal to be reconstructed, we propose to recover the motion-compensated residual, rather than the original field, using a TV-based reconstruction method.

Experiments show that the proposed algorithm generates high quality results, having an average of 5dBs PSNR gain compared to other deinterlacing approaches.

References

1. Kim, W., Jin, S., Jeong, J.: Novel intra deinterlacing algorithm using content adaptive interpolation. IEEE Trans. Consum. Electron. 53(3), 1036–1043 (2007)
2. Biswas, M., Kumar, S., Nguyen, T.Q.: Performance analysis of motion-compensated deinterlacing systems. IEEE Transactions on Image Processing 15(9), 2596–2609 (2006)
3. Mohammadi, H.M., Langlois, P., Savaria, Y.: A five-field motion compensated deinterlacing method based on vertical motion. IEEE Trans. Consum. Electron. 53(3), 1117–1124 (2007)
4. Fan, Y.C., Lin, H.S., Chiang, A., Tsao, H.W., Kuo, C.C.: Motion compensated deinterlacing with efficient artifact detection for digital television displays. Journal of Display Technology 4(2), 218–228 (2008)
5. Chen, Y.R., Tai, S.C.: True motion-compensated de-interlacing algorithm. IEEE Transactions on Circuits and Systems for Video Technology 19(10), 1489–1498 (2009)

6. Trocan, M., Mikovicova, B., Zhanguzin, D.: An adaptive motion-compensated approach for video deinterlacing. Springer's International Journal on Multimedia Tools and Applications, 1–19 (July 2011), doi:10.1007/s11042-011-0845-7
7. Trocan, M., Mikovicova, B.: Smooth motion-compensated video deinterlacing. In: Proc. of 7th International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 143–148 (September 2011)
8. Keller, S., Lauze, F., Nielsen, M.: A Total Variation Motion Adaptive Deinterlacing Scheme. In: Kimmel, R., Sochen, N.A., Weickert, J. (eds.) Scale-Space 2005. LNCS, vol. 3459, pp. 408–418. Springer, Heidelberg (2005)
9. Keller, S., Lauze, F., Nielsen, M.: An adaptive motion-compensated approach for video deinterlacing. IEEE Transactions on Image Processing 17(11), 2015–2028 (2008)
10. Yin, X., Yuan, J., Lu, X., Zou, M.Y.: De-interlacing technique based on total variation with spatial-temporal smoothness constraint. In: Science in China Series F: Information Sciences, vol. 50, pp. 561–575 (2007), doi:10.1007/s11432-007-0047-0
11. Li, C.: An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing, M.S. thesis, Rice University (September 2009)
12. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Transactions on Signal Processing 57(7), 2479–2493 (2009)
13. Chan, T.F., Esedoglu, S., Park, F., Yip, A.: Total variation image reconstruction: Overview and recent developments. In: Paragios, N., Chen, Y., Faugeras, O.D. (eds.) Handbook of Mathematical Models in Computer Vision, vol. 2, Springer, New York (2006)
14. Bioucas-Dias, J.M., Figueiredo, M.A.T.: A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. IEEE Transactions on Image Processing 16(12), 2992–3004 (2007)
15. Wang, Y., Yang, J., Yin, W., Zhang, Y.: A new alternating minimization algorithm for total variation image reconstruction. SIAM Journal on Imaging Sciences 1(3), 248–272 (2008)
16. Trocan, M., Maugey, T., Tramel, E.W., Fowler, J.E., Pesquet-Popescu, B.: Compressed sensing of multiview images using disparity compensation. In: Proceedings of the International Conference on Image Processing, Hong Kong, pp. 3345–3348 (September 2010)
17. Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis, Ph.D. thesis, Massachusetts Institute of Technology (May 2009)

Chapter 4
Cartographic Representation of Route Reconstruction Results in Video Surveillance System

Karol Lisowski and Andrzej Czyżewski

Abstract. The video streams available in a surveillance system distributed on the wide area may be accompanied by metadata are obtained as a result of video processing. Many algorithms applied to surveillance systems, e.g. event detection or object tracking, are strictly connected with localization of the object and reconstruction of its route. Drawing related information on a plan of a building or on a map of the city can facilitate the perception of events. Methods of augmenting cartographic data are proposed in this chapter. Making it possible to merge and to present a large amount of useful data on a single screen of surveillance.

4.1 Introduction

Concerns about the visual perception of multiple video images as well as problems with concentration on more than one field of view for a long time are widely discussed in the literature[1, 2, 4]. The user of a multi-camera surveillance network can acquire knowledge about spatio-temporal dependencies between cameras while using this system. Therefore the user can predict location and time in which a particular object appears again in another camera. Thus the whole video surveillance system can be described as a graph. The graph can be put on the map of the monitored area and hence locations seen by cameras are linked with places on the map. The user can get a synthesized view on the supervised area in the graphical form. In this case names of video streams do not have to be named as locations of placement, because this information is available on the augmented map at first glance. In addition, if a video surveillance system detects a danger in some places, those events can be linked together on the map. The paths and events connected with a particular object can be drawn on the surface of map. Methods for prediction of movement

Karol Lisowski · Andrzej Czyżewski
Multimedia Systems Department, Gdańsk University of Technology,
Narutowicza 11/ 12, 80-952 Gdańsk, Poland
e-mail: {lisowski,andcz}@sound.eti.pg.gda.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 35–44.
springerlink.com

can be implemented as well. Metadata from analysis of video streams, like markers and blobs connected with particular object, can be put on the video image and detected events can be presented as text. Concatenation of these fragmented data can be made only by an experienced user who exactly knows the topology of the camera network. Putting the topology on the map, as a graph, and presenting metadata related to each camera as markers makes it possible to see the whole context of activity of a particular object.

4.2 Related Work

Considerable effort has been put on developing methods for getting contextual meaning of video streams from surveillance system. In case of non-overlapping FOVs (*Fields Of Views*) re-identification methods are used to track the movement of objects continuously between cameras. A solution for this problem include analysis of color of object and matching histograms [3]. Working in a continuously changing environment, where light changes and differences in white balance can occur, causes necessity of compensation of color or making color descriptors independent to these obstructions[5, 10, 9]. Tracking within one camera also must be implemented in order to get information about appearing and disappearing of the object. Moreover obtaining visual features of object is important for re-identification methods. Additionally, the topology of placement of the cameras is needed. This information can be obtained automatically as a result of topology recovery algorithms[11, 8, 12, 17]. In order to enrich the map with more semantic data, algorithms for event detection and object classification need to be implemented in the video surveillance system[15, 7, 6]. Tracking of objects is presented within FOV of the camera as markers but transitions between cameras and events can be described in a textual form only. To fill in this gap the method for augmenting cartographic data can be implemented.

4.3 Proposed Method

4.3.1 Topology of Camera Network

The first step to put data from analysis of video streams on the map is describing the topology in the form of graph G. The graph is composed of three parts i.e.: set of nodes N, set of edges E and parameters P assigned to each edge.

$$G = \{N, E, P\} \tag{4.1}$$

Each node of the graph is related to a particular camera and determines its location. Likewise, each edge relates to possible transitions between FOVs of cameras. The parameters are used to describe attributes of edges and thereby transitions between cameras, as well. Generally graph parameters can be considered as a matrix in which each element describes unidirectional transition between particular pair of nodes. Thus, e.g. an element p_{34} contains parameters of transition from node 3 to node 4:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \tag{4.2}$$

An example attribute of the transitions is the maximum period of time or probability density function of the time of transition. The graph of topology can be obtained through the analysis of placement of the cameras by a human, which is possible for small systems, but also automatic algorithms and methods can be used. Putting the graph on the map or plan demands assigning real coordinates of camera placement for all nodes.

4.3.2 Obtaining Contextual Data

Object Detection: The first step of analysis of video data is making distinction between moving objects and static background. Methods for background subtraction and for object detection must be implemented. These methods are based on GMM (*Gaussian Mixture Model*) or Codebook algorithm as described in more details in literature[13, 7]. Regions in image determined as objects are the result of object the detection algorithm execution.

Object Classification: When objects in video image are detected, appropriate regions of interest are analyzed in order to determine types (classes) of given objects. The classification occurs on the basis of the size of object in video image and its speed of movement. Related algorithms are described in the literature[7, 6, 15].

Object Tracking: The results of object detection are also used for tracking objects within video image. Assigning trackers to objects and moving them occurs owing to algorithms based on Kalman filters. Those filters use descriptors of color and texture in order to estimate movement of particular trackers. Related details of the applied object tracking method can be found in the literature[16, 7].

Event Detection: Data from object detection and classification can be used for analysis of higher contextual level events. Example events like entrance into forbidden area, luggage abandonment or other suspicious behavior can be detected on the basis of previously defined rules. Details are contained in [15, 7, 14].

Tracking between cameras: As mentioned above locations of cameras in a video surveillance system can be described as vertices of topology graph. While movement of the object between cameras FOVs proceeds cameras exchange messages in order to build path of each object within whole system. The message contains visual features descriptors, timeout (determining time period in which given object is expected by recipient) and timestamp of object disappearing from FOV of camera sending this message. The group of recipients (the neighborhood of recipient) and timeouts are taken from the topology graph. Each of cameras has a table of past messages and for those generated whenever object appears in any FOV the comparison and matching occurs. In case of identifying the same object the next node is added to the path of this object and information about a match is sent to remaining recipients and to the sender of the message. When the match didn't occur during whole timeout period, the message is removed from the table. The flowchart of messaging between cameras in an example situation is presented in Fig.4.1. The path of *object X* is marked by the red arrows. An *object X* appeared in *Camera A* and was not found in its table, thus it is recognized as a brand-new object in the system. When *object X* left *camera A* messages about it are sent to the neighborhood of this node in the topology graph (that is *Camera B* and *Camera C*). Subsequently, *object X* left *Camera B* and made its way towards *Camera D* where it was also identified as *object X*. After being seen in *Camera D* object left the system. Despite the fact of sending messages to the neighborhood of *camera D* no matching was made, thus these messages are afterwards, removed because of the timeout constrains.

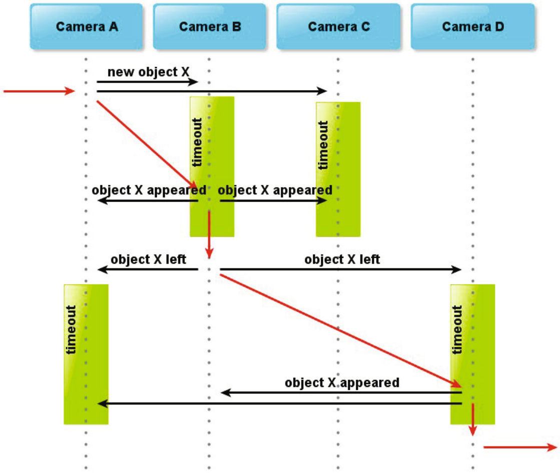


Fig. 4.1 Signals between cameras while object movement occurs

As a result of video processing the metadata from all cameras are sent to server and stored there. A user connected to the server can browse history of movement available on this server and is able to receive data about actual activity of objects in real-time.

4.3.3 Architecture of the System

Certain structure of video surveillance system should be utilized in order to implement augmented map of the supervised area. The whole video processing can be carried out within nodes and then results are collected by the server. The server arranges data into groups related to particular nodes. The ordered data are sent to the user where they can be linked to cartographic images. The linkage occurs on the basis of the topology graph which was previously put on the map. The architecture of the related system is presented in Fig.4.2. Such an architecture allows for sending data from many cameras to the one place where the data is processed and forwarded to the users.

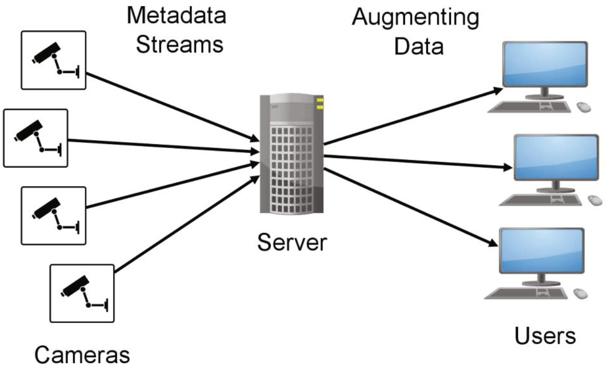


Fig. 4.2 Architecture of the system

4.3.4 Augmenting Cartographic Data

The video surveillance system generates different sorts of data as results of various methods of video processing used to obtain contextual meaning of observed situations. Those methods include:

- **tracking objects:** two levels of tracking can be considered, that is within single FOV and between pairs of cameras. For augmenting cartographic data determination of the transitions between FOVs of cameras is most important. Object tracking provides spatio-temporal context related to the appearance of a given object. Therefore, by means of re-identification methods implemented to the server, the movement path for this object can be recovered. If a pair of observations is classified as a match a part of the path can be put on the map or on the plan. Each tracked object gets its unique identifier which follows the object as it keeps moving. Thus, augmented data from object tracking objects are grouped by unique identifiers and they contain locations and times of observations:

$$o_i \in O = \{(l,t,i) : l \in L, t \in T, i \in I\} \tag{4.3}$$

o_i is the observation related to the object with assigned identifier i. The values of l, t and i represent components of the observation, whereas L, T and I are sets of all localizations of cameras, certain period of time and all assigned identifiers, respectively.

- **classifying objects:** the output of the classification can be also presented on the map. Most frequently the classification is performed on the basis of the appearance of the object, but also such parameters as speed of movement within FOV can provide an additional information in this process. The result of classification for a given object is assigned to its unique identifier:

$$c_i = \{c \in C\} \tag{4.4}$$

  where c_i is a class of the object with assigned identifier i and C is the set of classes used to classify.
- **event detection** gives more semantic information than previous methods but in the same time it is based on them. The outcome of the event detection is determination if any of rules, that is connected with the defined type and location of event, are fulfilled. Then data about the event can be sent to the augmented map. This operation is described by Eq.4.5.

$$e_{ts} \in E = \{(s,t,l) : s \in S, t \in T, l \in L\} \tag{4.5}$$

  where E is the set of events, S is set of types of events.

As each object gets identifier, thus markers described by this identifier can be put on the map. The class of object and events can be assigned to the marker, as well.

4.4 Results

In order to realize the idea presented above an application and GUI (Graphical User Interface) were prepared. The application was created for the operator (user) usage as a front-end of working system which architecture was shown in Fig. 4.2. Moreover, methods listed and shortly described in Sec. 4.3.2 were used on consecutive levels of video analysis so that metadata for user application are gathered on the server and can be transformed to the augmenting data. Graphical user interface, which is created in QT framework, provides features to place and move camera markers on the map manually with mouse using drag-and-drop action. Linking the camera markers is nothing else than building the topology graph which is used as the frame for the drawing augmenting data on the map. The user has also possibility to send created topology graph to the server where it will be used in Route Reconstruction algorithms. The screenshot presenting GUI of prepared application is shown in Fig. 4.3.

The basis for the perceptual augmenting process is a map or a plan of the monitored area. It is an image in which the topology of camera network was marked, so that subsequently data from video surveillance system can be used for augmenting the cartographic data. As it was mentioned above there is a necessity of adding the topology graph feature. The nodes and edges can be set manually or by providing

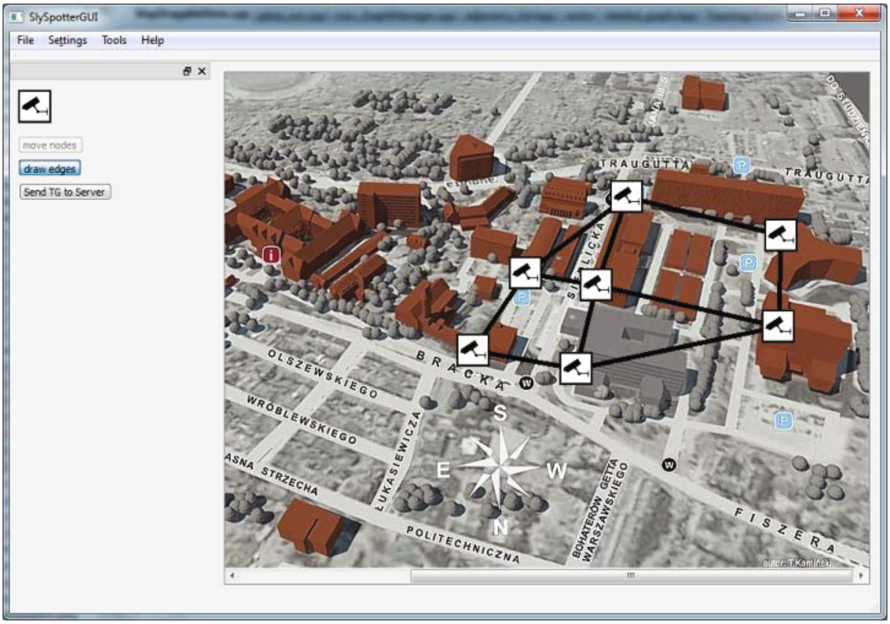


Fig. 4.3 GUI of prepared application containing topology graph of video surveillance network put on the plan

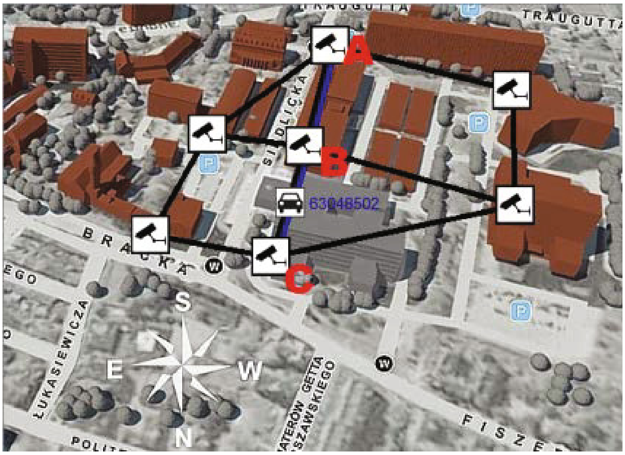
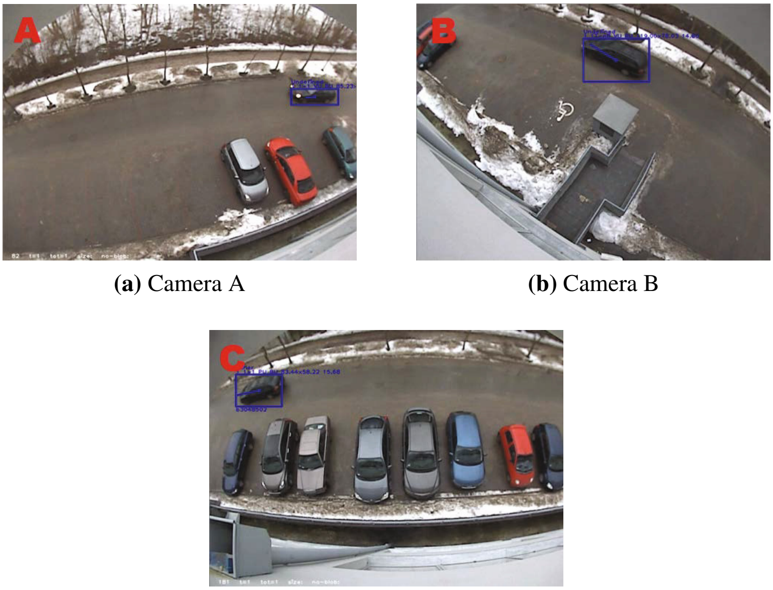


Fig. 4.4 Map augmented with information about movement (blue line) and type of object

coordinates of cameras. The result is presented in Fig.4.3 where possible transitions are depicted.

In the situation presented in Fig.4.4 an object, which is classified as a car, was moving along the road. From the moment of entering FOV of camera A the object obtains an identifier which is assigned to it until disappearing from FOV of the last camera on the object's path. The blue lines, which are tangential with edges of the graph, present the path of object's movement through the camera network. Consecutive parts of the path are drawn on the map as the object changes its position.

The video surveillance system works to some extent in the background of augmented map, thus nothing prevents viewing video images from particular cameras. In Fig.4.5 video images from cameras which record the movement of object are presented. Thus this kind of "live map" can be the main part of the control panel for a video monitoring system, because of providing the context of video image from each camera.



(a) Camera A

(b) Camera B

(c) Camera C

Fig. 4.5 Video images recorded from path of tracked object which path was presented in Fig. 4.4

4.5 Conclusion

The usage of augmented map enables an integrated view on the whole video surveillance system. It does not offers as much details as in each video image does, however the context of the observed situation is usually easier to understand. The operator gets data from the analysis of video in many cameras which are filtered and prepared in a graphical form. Putting the graph of topology on the map helps to understand events which are taking place in video images.

The next outcome is the possibility of viewing the video images by selecting them on the map. Such a functionality may add a spatio-temporal context to the video stream preceding its retrieval and watching.

Acknowledgements. This work is associated with the ADDPRIV ("*Automatic Data relevancy Discrimination for a PRIVacy-sensitive video surveillance*") Project No. 261653, being a part of the European Seventh Framework Program (FP7).

References

[1] Allen, R., Mcgeorge, P., Pearson, D., Milne, A.B.: Attention and expertise in multiple target tracking. Applied Cognitive Psychology (2004)

[2] Arthur, F., Kramer, S.H.: Splitting the beam: Distribution of attention over noncontiguous regions of the visual field. Psychological Science (1995)

[3] Cai, Y., Chen, W., Huang, K., Tan, T.: Continuously Tracking Objects Across Multiple Widely Separated Cameras. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part I. LNCS, vol. 4843, pp. 843–852. Springer, Heidelberg (2007)

[4] Cavanagh, P., Alvarez, G.: Tracking multiple targets with multifocal attention. Trends in Cognitive Sciences (2005)

[5] Colombo, A., Orwell, J., Velastin, S.: Colour constancy techniques for re-recognition of pedestrians from multiple surveillance cameras. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (2008)

[6] Dalka, P., Czyżewski, A.: Vehicle Classification Based on Soft Computing Algorithms. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 70–79. Springer, Heidelberg (2010)

[7] Dalka, P., Szwoch, G., Szczuko, P., Czyżewski, A.: Video content analysis in the urban area telemonitoring system. In: Multimedia Services in Inteligent Environments. Springer, Heidelberg (2010)

[8] Farrell, R., Davis, L.S.: Decentralized discovery of camera network topology. IEEE (2008)

[9] Gilbert, A., Bowden, R.: Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In: 9th European Conference on Computer Vision, ICCV 2006 (2006)

[10] Javed, O.: Appearance modeling for tracking in multiple non-overlapping cameras. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 26–33 (2005)

[11] Nama, Y., Ryu, J., Choi, Y., Cho, W.: Learning spatio-temporal topology of a multi-camera network by tracking multiple people. World Academy of Science - Engieneering and Technology (2007)

[12] Niu, C., Grimson, E.: Recovering non-overlapping network topology using far-field vehicle tracking data. In: The 18th International Conference on Pattern Recognition, ICPR 2006 (2006)

[13] Szwoch, G.: Performance Evaluation of the Parallel Codebook Algorithm for Background Subtraction in Video Stream. In: Dziech, A., Czyżewski, A. (eds.) MCSS 2011. CCIS, vol. 149, pp. 149–157. Springer, Heidelberg (2011)

[14] Szwoch, G., Dalka, P.: Automatic detection of abandoned luggage employing a dual camera system. In: MCSS 2010: IEEE International Conference on Multimedia Communications, Services and Security (2010)

[15] Szwoch, G., Dalka, P., Czyżewski, A.: Objects classification based on their physical sizes for detection of events in camera images. In: NTAV/SPA 2008 Signal Processing: Algorithms, Architectures, Arrangements, and Applications; New Trends in Audio and Video, pp. 15–20 (2008)

[16] Szwoch, G., Dalka, P., Czyżewski, A.: Resolving conflicts in object tracking for automatic detection of events in video. Elektronika (2011)

[17] Tieu, K., Dalley, G., Grimson, W.E.L.: Inference of non-overlapping camera network topology by measuring statistical dependence. In: Proceedings of the Tenth IEEE International Conference on Computer Vision, ICCV 2005 (2005)

Chapter 5
Visual Objects Description for Their Re-identification in Multi-Camera Systems

Damian Ellwart and Andrzej Czyżewski

Abstract. The topic of object tracking in video surveillance systems, is addressed in this work. An introduction to techniques of single camera and multi camera object tracking is presented. The problem of robust visual object description is discussed. Implemented image parameterization methods, including algorithms based on the MPEG-7 standard, are shown. Examples of the prepared dataset from a multi-camera system are presented. Chosen descriptors evaluation, employing this dataset, is performed. Descriptors evaluation procedure is described in detail. The results utilizing distance measures are compared. Conclusions based on performed experiments are described. Scope of the future work is outlined.

5.1 Introduction

Image analysis can be applied practically to many task, like: traffic signs recognition system, tools for video indexing and browsing, robotics vision, web contents indexing, product quality checks, face recognition, aids for visually impaired people, medical diagnostics and many more. The main purpose of utilizing many of those systems are related to security reasons. Especially, large video monitoring systems, are difficult to manage. Therefore, this work presents authors contribution to smart surveillance systems, in the context of multi camera image analysis.

Smart video surveillance systems can be divided into three general groups according to their architecture.

In the first system type, camera recordings are streamed to a dedicated and adequately effective device, which performs image analysis. In the second architecture type, dedicated devices are utilized for each camera in the system. Those

Damian Ellwart · Andrzej Czyzewski
Gdansk University of Technology
Narutowicza 11/12, 80-233 Gdansk, Poland
e-mail: {ellwart,andcz}@sound.eti.pg.gda.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 45–54.
springerlink.com

devices produce metadata describing video contents which are further sent to a centralized unit. Last system type, benefits from so called smart cameras, which can carry out image processing for the purpose of event detection. In a way, the second and third types are similar but the "black box" devices presented in the second approach are capable to be integrated within the camera enclosure.

Regardless of the hardware solution, similar image processing techniques are typically utilized. Before approaching to event detection, a number of analysis stages are required to be performed. Depending on the problem and the environment, chosen algorithms can vary but a general logical data flow within the system can be depicted as in Fig. 5.1.1.

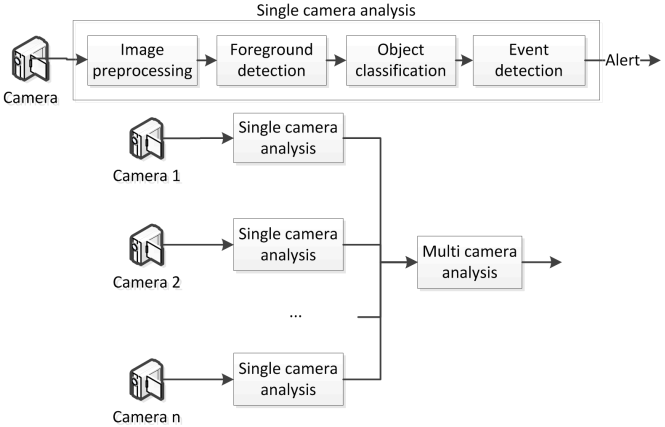


Fig. 5.1.1 Smart video surveillance system image processing scheme example

As it is shown in Fig. 5.1.1, two analysis phases can be distinguished. First, is related to single camera processing, where each video is treated independently. At this phase, foreground extraction and its analysis can be performed leading to defined events or abnormalities detection. The second stage is devoted to a combined analysis of single camera processing results. Here, the problem of object re-identification can be introduced. The purpose of such an analysis is to identify the same objects (i.e. people) which can be variously represented in different system cameras. This process may allow object tracking to be performed, within the whole monitored area. A complete route of a moving object can be obtained on demand, in this way. However, this task requires several image processing steps to be performed. In the following section several approaches for single- and multi-camera object tracking are presented. Completing of above algorithmic steps are essential for a smart surveillance system to achieve the aforementioned goal.

5.2 Object Tracking

Object detection process, regardless of the utilized method (i.e. background removal), produces information about moving objects present in the analyzed scene. This action is performed for all consecutively analyzed video frames. The detected objects are typically processed further to acquire detailed information. To make the analysis more consistent, object relations between analyzed frames but also between different cameras can be introduced. This is the main goal of the object tracking module existence within a smart surveillance system. In the following subsections some popular techniques for objects detection are reviewed briefly.

5.2.1 Single Camera Object Tracking

The task of single camera object tracking is to introduce relations between objects in consecutive video frames. To perform this continuously, several problems need to be overcome. This process needs to be robust against object detection inaccuracy and temporal partial and full occlusions which can occur when an object moves behind a scene obstacle (i.e. pole, parked vehicle, tree). To solve this problem, commonly, Kalman filtering based approach is used [1, 2]. In this way, each object location and dimensions are described by a state, which is updated accordingly to new object observations. Additionally, it is possible to estimate the predicted position and size of an object. This information can be used in case of object occlusion.

Unfortunately, in real life scenarios the description and prediction of just objects paths is not sufficient for continuous object tracking to be satisfied. Therefore additional tracker assignment rules are needed. In literature [3] the method called Blob Matching is used where every object is described by a set of features, corresponding to its appearance. Detected objects feature vectors in each processed frame are matched with the available trackers or a new tracker instance is created. Still, this approach fails often when similar objects are present in the scene. A combination of Kalman filtering and feature extraction and matching techniques overcome this problem partially as the object visual description is supported by its path estimation and vice versa [4]. Robust object tracking is an essential part of a video processing system. It allows object continuous analysis to be done during its presence in the monitored area.

5.2.2 Multi-Camera Object Tracking

Object tracking process can be applied to a multi-camera system as well. This way, the analyzed objects are tracked not only within a single scene but in every camera working in the system. Such a tracking makes a difficult task, since objects appearance in independent cameras may vary a lot. However, it can provide relevant information about the mobility statistics and the most frequently used paths under surveillance. Moreover, if the object of interest is tagged (manually by the user or automatically via an event detector), it might be possible to reconstruct its whole route.

To find an object which was tagged in one of the system cameras, other video sources need to be verified for this object presence. Provided consecutive system cameras views are overlapping, direct relations between observed objects can be found [5] making the tracking process relatively simple. If this is not the case, then multi-camera object tracking becomes difficult and can be less effective. However the accuracy of this process can be increased by taking into account the topology of the system. Consequently, the object of interest presence needs to be verified in some selected cameras. That means also that only cameras adequately spatially related to the camera, where the object was visible last, are considered. Additionally, the process can be aided by applying a specific time window for each of the cameras during the search for the tracked object. The size of the time window typically varies depending on the system topology and is estimated upon observations.

Regardless of the mentioned assumptions, a proper method for object recognition in multi-camera system needs to be utilized. There are several ways to achieve this goal [6]. However, the basic idea for object re-identification involves proper visual object description. In the following section, the problem of robust feature extraction in multi-camera system is discussed. Afterwards, chosen description techniques are presented and evaluated.

5.3 Object Description Methods

In real life conditions, cameras working within surveillance systems are rarely identical. For this reason images acquired from each of them may be different, even if their fields of view cover the same area. This dissimilarity can be caused by various camera parameters like exposure, white balance and many more, but also can result from different camera matrix characteristics. In addition to that, the object appearance can be changed by a variety of lighting conditions. In case of outdoor monitoring, shadowed areas can make the tracked object unrecognizable. Similar problems can be introduced, when the area under surveillance is a mixture of outdoor and indoor scenes. In such a case, the shadowed areas are only one of the inconveniences. Indoor lighting conditions can vary as well, depending on the applied illumination type. The presented situations cause major difficulties which have to be dealt with, during the re-identification module development.

There is a large variety of visual object descriptors which can be found in literature [7-9], but only a part of them are typically applied to the considered problem. As it was already stated, object description for object re-identification in a multi-camera surveillance system needs to be robust against scene illumination changes and the differences resulting from image acquisition using various camera types. Therefore, the extracted set of features needs to be invariant to a set of transformations [8]. Most popularly utilized descriptors for this problem include SIFT and color histograms in various color spaces and representations [10-14]. SIFT based description offers good results in most of the image recognition problems [15], however its calculation process is quite complex (even considering SURF image descriptor [7]). Additionally in particular cases, especially for low resolution images, the acquired description can be poor due to the difficulties occurring during key-point detection stage.

Nevertheless, many more feature extraction techniques were introduced over the years in the field of image recognition [9, 16-18]. Several of them were implemented to verify their usefulness in the object re-identification problem as an alternative to local features [7]. The chosen descriptors include:

- Color histogram (RGB, Transofmed color)
- Color Layout Descriptor (YCrCb, Transformed color)
- Color Moment Invariants GPSO (RGB)
- Local Binary Pattern histogram (RGB)
- Edge Histogram Descriptor (Greyscale)

Color histogram in the RGB color space is the most basic method of representing visual content. Still it is utilized as a reference to other description techniques. To improve this representation robustness, image histogram for Transformed color was built as such representation is known to be invariant to offsets, illumination intensity and light color changes [8]. Besides characterizing the basic color pallet present in the image, additional information about its spatial distribution can be important. Hence, Color Layout Descriptor was implemented for two color spaces, according to the procedure presented in [19, 20]. However, for this method additional preprocessing was required since CLD is defined for images with regular dimensions. During this stage the described objects are oriented vertically and stretched in each row independently to fit the object bounding rectangle. Another implemented method utilizes a set of image moment-based parameters proposed in literature which are meant to be invariant to a set of image transformations [18]. They are called Moment Invariants GPSO and are expressed as a combination 2 and 3 band image moments. Last two descriptors are related to the information on the object texture. First of them forms a histogram from the binary words extracted on the basis of the Local Binary Pattern transformation [21, 22]. This method describes spatial relations between neighboring image points. Hence, it introduces a level of invariance to global lighting changes. The second texture description technique depicts distribution of image edges [16, 19]. This method requires image preprocessing procedure, same as for CLD, to be performed.

To evaluate the aforementioned visual object description techniques, appropriate experiments were carried out. This process is presented in detail in the following section.

5.4 Experiments

As it is stated in the previous sections, multi-camera object tracking requires a proper object description to be used. To evaluate implemented parameterization methods a number of recordings was prepared. The videos were acquired using a multi-camera setup monitoring a hall and a set of corridors leading to it. Various people could be observed by different cameras in above conditions. The camera devices used, were directed to the scene at similar angles. Each camera captured a different part of the monitored area. Moreover, field of view-related illumination differences between cameras were noticed. This effect could be at least partially compensated utilizing the Brightness Transfer Function approach [23]. However, this work is focuses on parameterization invariance analysis under typical surveillance conditions and therefore camera color calibration is not taken under considerations. Hence, the illumination variety can result in object re-identification difficulties which were described in the multi-camera object tracking section. Sample frames from the acquired videos are presented in Fig. 5.4.1.

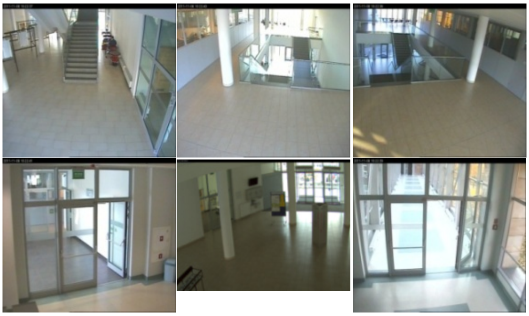


Fig. 5.4.1 Sample video frames from the prepared recordings

The recordings were obtained from six video sources, including two types of cameras with different characteristics. The material was captured with two resolutions, depending on camera type: 704x628 and 1920x1080.

5.4.1 Dataset Preparation

From the acquired recordings, 160 images of objects were manually extracted using editing tools. In this way, descriptors examination becomes independent from any errors which could occur on the previous processing stages (i.e. foreground detection). The acquired image set shown in Fig. 5.4.2 represents 11 objects in 6 different cameras.

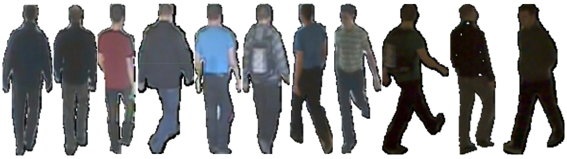


Fig. 5.4.2 Visual object examples extracted from prepared recordings

For the purpose of evaluation, the obtained gallery of images was organized hierarchically. The images representing the same people were grouped. This is useful to determine the separation between different object description techniques within particular cameras. Additionally, the object examples are sorted according to the camera used to record them. That was done, in order to ensure the stability of the representation of images recorded by cameras. Further experiments treat both of these cases independently in the first step. Afterwards, the results are compared against each other to depict the overall quality of the produced object description.

5.4.2 Feature Evaluation

On the basis of implemented descriptors and the prepared dataset, feature vectors were generated. To compare the descriptors, inner and outer-class scatter is verified. As an example, the results for one of the implemented descriptors is presented in Table 5.4.1. To assess the compactness of each class, mean and standard deviation of the distance from cluster center is calculated (Table 5.4.1, right part). On the other hand, to depict classes separation, cluster center distances are utilized (Table 5.4.1, left part). Additionally, to test whether the classes are disjunctive the relationship expressed by eq. 5.4.1 is checked for each pair, independently.

$$\mathrm{sep}_{x_a x_b} = \begin{cases} \mathrm{true}, \mathrm{d}(\overline{x_a}, \overline{x_b}) > \sigma(x_a) + \sigma(x_b) \\ \mathrm{false}, \mathrm{d}(\overline{x_a}, \overline{x_b}) < \sigma(x_a) + \sigma(x_b) \end{cases} \tag{5.4.1}$$

where $\mathrm{d}(\overline{x_a}, \overline{x_b})$represents the Euclidean distance between class centers and σ(x) is the class standard deviation both calculated along for analysed clusters.

Table 5.4.1 Description results for color histogram in Transformed color space

Obj. id	1	2	3	4	5	6	7	8	9	10	11	mean	std
1	0	0.913	0.790	1.143	0.750	1.301	1.087	1,319	1,258	1,213	1,525	0,304	0,371
2	0,913	0	0.717	0,883	0,930	1,305	1,008	1,069	0,986	1,295	1,721	0,345	0,317
3	0,790	0,717	0	0,774	0,941	0,796	0,836	1,196	1,266	1,055	1,246	0.416	0.410
4	1.143	0.883	0.774	0	1.335	0.928	1.113	1.461	1.245	1.264	1.440	0.318	0.211
5	0.750	0.930	0.941	1.335	0	1.549	1.387	1.048	1.215	1.424	1.877	0.231	0.215
6	1.301	1.305	0.796	0.928	1.549	0	0.910	1.640	1.652	1.095	1.024	0.424	0.192
7	1.087	1.008	0.836	1.113	1.387	0.910	0	1.390	1.286	1.036	1.250	0.454	0.282
8	1.319	1.069	1.196	1.461	1.048	1.640	1.390	0	1.331	1.464	2.063	0.138	0.016
9	1.258	0.986	1.266	1.245	1.215	1.652	1.286	1.331	0	1.484	2.071	0.107	0.018
10	1.213	1.295	1.055	1.264	1.424	1.095	1.036	1.464	1.484	0	1.346	0.192	0.019
11	1.525	1.721	1.246	1.440	1.877	1.024	1.250	2.063	2.071	1.346	0	0.157	0.008

The darkened cells in the resulting table, reveals that the condition expressed by eq. 5.4.1 is not fulfilled. This means that the classes are not fully separable for the corresponding objects. Similar analysis was carried out for each of the feature extraction methods. The experiments, show that most descriptors are sufficient to distinguish objects employing a single camera. However, while analyzing the results, it can be noticed that for the tested dataset, some of the parameterization techniques are not sufficient to recognize objects representations from various cameras, while utilizing distance measures. For most of the descriptors, the most scattered clusters are related to objects with ids: 1, 3 and 7. These results are in accordance with objects' actual appearance since their front and back views differ most significantly in the prepared representations. Sample images of these objects are illustrated in Fig. 5.4.3.

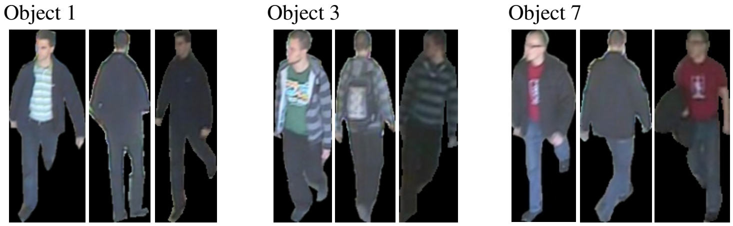


Fig. 5.4.3 Representations of objects with the most diverse appearance

Summary of the results is presented in Table 5.4.2. It shows rates related to fully separable cluster pairs for each of the descriptors, with an additional information about the parameter vector length.

The score of 100% in Table 5.4.2 indicates that the clusters, consisting of objects feature vectors, are disjoint. From the analyzed description methods LBP histogram and both CLD types represented the dataset most robustly. On the other hand, Moment Invariant GPSO descriptor is worst. It may be the result of a short feature vector utilized in this case which is not sufficient to describe the visual object properly. Regular RGB color histogram turned out to perform poor. The reason for that can be related with significant object representation differences in various cameras. Similar separability results can be observed for Edge Histogram Descriptor. This result might be expected, since a typical person outfit does not contain a lot of edges and high contrasts.

Table 5.4.2 Feature extraction methods result summary

Descriptor	Separated classes pairs	Feature vector size
Color Histogram	53% (29/55)	192
Color Histogram Trans.	85% (47/55)	192
CLD	100% (55/55)	192
CLD Trans.	85% (47/55)	192
LBP Histogram	95% (52/55)	192
EHD	51% (28/55)	60
Moment Invariants GPSO	45% (25/55)	18

5.5 Summary

An introduction to single and multi-camera tracking topic was presented in this document. Typical problems and possible solutions for this task were mentioned. Further, a short overview of feature extraction techniques was presented. Some of the chosen methods including two, based on MPEG-7 standard, were described in detail. In the last section experimental procedure was shown and the results discussion was carried out. It was shown that other than currently extensively utilized SIFT-based descriptors can by successfully used for object re-identification in disjoint camera views. For the prepared dataset CLD and LBP turned out to be a good choice.

Further work will include classifier employment (i.e. ANN, SVN, Decision Trees) as well as more broad experiments involving a larger dataset of objects. Additional descriptors will be evaluated utilizing more sophisticated cluster segmentation validity measures.

Acknowledgments. Research subsidized from the budget of the project ADDPRIV ("Automatic Data relevancy Discrimination for a PRIVacy-sensitive video surveillance") No. 261653, being a part of the European Seventh Framework Program.

References

[1] Czyżewski, A., Dalka, P.: Examining Kalman filters applied to tracking objects in motion. In: Proc. of 9th International Workshop on Image Analysis for Multimedia Interactive Services, pp. 175–178 (2008)

[2] Kim, Y.M.: Object Tracking in a Video Sequence. Final Project Report (2006)

[3] Fuentes, L.M., Velastin, S.A.: People tracking in surveillance applications. Image and Vision Computing (2006)

[4] Han, Z., Ye, Q., Jiao, J.: Online feature evaluation for object tracking using Kalman filter. In: 19th International Conference on Pattern Recognition, pp. 1–4 (2008) ISBN: 978-1-4244-2174-9

[5] Yilmaz, A., Javed, O., Shah, M.: Object tracking–a survey. ACM Computing Surveys (2006), doi:10.1145/1177352.1177355

[6] Prosser, B., Gong, S., Xiang, T.: Multi-camera Matching under Illumination Change Over Time. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications, pp. 1–12 (2008)

[7] Bay, H., Tuytelaars, T., Gool, L.V.: SURF: Speeded Up Robust Features. Computer Vision and Image Understanding 110, 346–359 (2008)

[8] Sande, K.E., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1582–1596 (2010)

[9] Pinheiro, A.M.G.: Image Descriptors Based on the Edge Orientation. In: 4th International Workshop on Semantic Media Adaptation and Personalization, pp. 73–78 (2009)

[10] Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. Distributed Smart Cameras, 1–6 (2008)

[11] Teixeira, L.F., Corte-Real, L.: Video object matching across multiple independent views using local descriptors and adaptive learning. Pattern Recognition Letters, 157–167 (2008)

[12] Quelhas, P.: Natural scene image modeling using color and texture visterms. Image and Video Retrieval (2006)

[13] Piccardi, M.: Multi-frame moving object track matching based on an incremental major color spectrum histogram matching algorithm. Computer Vision and Pattern 3, 19–24 (2005)

[14] Cai, Y., Huang, K., Tan, T.: Human appearance matching across multiple non-overlapping cameras. In: 19th International Conference on Pattern Recognition (2008)

[15] Tao, Y., Skubic, M., Han, T., Xia, Y.: Performance Evaluation of SIFT-Based Descriptors for Object Recognition. In: IMECS, pp. 17–20 (2010)

[16] Balasubramani, R., Kannan, V.: Efficient use of MPEG-7 Color Layout and Edge Histogram Descriptors in CBIR Systems. Global Journal of Computer Science and Technology 9, 157–163 (2009)

[17] Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., Schmid, C.: Evaluation of GIST descriptors for web-scale image search. In: Proceeding of the ACM International Conference on Image and Video Retrieval (2009)

[18] Mindru, F., Tuytelaars, T., Gool, L.V., Moons, T.: Moment invariants for recognition under changing viewpoint and illumination. Computer Vision and Image Understanding, 3–27 (2004)

[19] Martínez, J.M.: MPEG-7 Overview (2004)

[20] Wei, Z., Zhao, R., Wu, Y., Zhu, J.: Efficient vehicle identification using MPEG-7 Color Layout Descriptor. In: International Conference on Business Management and Electronic Information, pp. 128–131 (2011)

[21] Topi, M., Matti, P., Timo, O.: Texture classification by multi-predicate local binary pattern operators. In: Proceedings on 15th International Conference on Pattern Recognition, vol. 3, pp. 939–942 (2000)

[22] Chang-yeon, J.: Face Detection using LBP features. Final Project Report, 1–4 (2008)

[23] Jeong, K., Jaynes, C.: Object matching in disjoint cameras using a color transfer approach. Machine Vision and Applications 19, 443–455 (2007)

Chapter 6
Compact Descriptor for Video Sequence Matching in the Context of Large Scale 3D Reconstruction

Roman Parys, Florian Liefers, and Andreas Schilling

Abstract. One of the key problems in the large scale reconstruction of 3D scenes from images is how to efficiently compute image relations in large databases. Finding images depicting the same 3D geometry is the pre-requisite for camera calibration and 3D reconstruction. In this chapter we present a simple and compact descriptor that enables us to efficiently compute similarity between video sequences. In addition to providing a similarity measure, the descriptor also makes it possible to select individual video frames that match together. With our descriptors, this computation can be done in a time similar to that required by the traditional SIFT algorithm to match just two images. Using the presented descriptors, we can build a large relation graph between video streams or image sequences. This relation graph is used later in assembling a large geometric model.

6.1 Introduction

In the field of computer vision, the topic of multi-view-stereo surface reconstruction has gained a lot of attention. As digital cameras become quite cheap and popular, there is a huge demand for algorithms, that are able to create models of 3D scenes directly from images. The competing approach - laser scanning - can deliver precise 3D models, but due to the high equipment prices, it is still out of reach for the consumer area. The state-of-the-art multi-view stereo methods are mature enough to deliver results, that can be compared to laser scans, in terms of accuracy [1].

The multi-view stereo algorithms are able to build 3D models using a set of images taken from different points of view, but nearly every algorithm requires to have a camera geometry provided. Camera parameters like focal length, radial distortion, position and orientation can be computed in many different ways, ranging from

Roman Parys · Florian Liefers · Andreas Schilling
Tuebingen University, Germany
e-mail: parys@gris.uni-tuebingen.de, florian@liefers.com,
schilling@uni-tuebingen.de

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 55–64.
springerlink.com

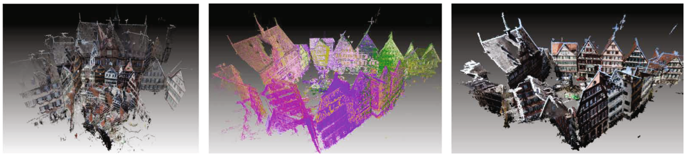


Fig. 6.1 Application of our compact descriptor to the problem of automatic merging of 3D models through creation of similarity graph. After creating a similarity graph between sequences of images used to reconstruct 3D sub-models in local coordinate systems (1st image), it was possible to merge the models (2nd and 3rd image) into a global reconstruction.

semi-manual methods that include known markers in a scene, to fully automatic methods. These methods need to have multiple views of the same scene in order to compute relations between cameras. Finding which images are observing the same scene is a challenging problem, from the perspective of computational complexity.

As the basic building block for our reconstruction algorithm are video sequences, the relations between images within sequences are known. The problem is how to compute image relations between different image sequences, as they can be taken at different times and by different people, while still showing the same geometry. Our contribution is the solution to this problem, by introducing a compact descriptor for video sequence matching. The descriptor enables us to evaluate similarity, and additionally to compute frame to frame relations between video sequences. The speed of our algorithm is comparable to the simple two frame matching algorithm presented in the classical SIFT paper [2]. The result from an example application of our compact descriptor for the automatic creation of a similarity graph and merging of 3D sub-models can be seen in Figure 6.1.

In this chapter we often use the term *image sequence*, however the algorithm does not assume any ordering of images within one subset. In the case of video sequences, we work on a subset of key frames with a good contrast and sufficient baseline.

6.2 Related Work

There have been many approaches used for finding similar video streams ([3], [4], [5], [6], and others), however in these cases, the frame ordering is a key factor. This does not fit into our scenario, where a sequence often is an unordered set of digital photographs mainly used for multi-view stereo reconstruction.

Another approach for finding similar videos, proposed in [7], is using an adaptive vocabulary tree to index all the video frames in the database, and each sequence is treated as a ”bag of frames”. The authors use global image features in order to reduce memory and computational requirements. This approach works well in context of copy detection, what is useful in detection of copyright infringements. In our scenario, global descriptors are not desired, as they often do not detect matching images useful for reconstruction. Additionally, the problem of matching videos is formulated as local alignment problem, which is not suited to our scenario.

In the rest of this section, we briefly describe the most influential work, that led to the derivation of our similarity descriptor and matching algorithm for image sequences, in context of 3D reconstruction.

One of the most important approaches for image matching is the Scale Invariant Feature Transform (SIFT) described in [2]. The author comes up with a selection of distinctive image features, that can be robustly matched with features of another image, under different scales and rotations.

There were approaches to optimize the speed of SIFT. The Speeded-up Robust Features (SURF) presented in [8] is aa successful attempt, that also has been implemented on Graphics Processing Units; however the penalty that has to be paid for the high speed of detection is robustness. Improvements to the matching speed of high dimensional vectors by casting them to a lower dimensionsional space were presented in [9].

the global scene appearance descriptor GIST, first described in [10] was used in [11], in a preliminary stage of similarity graph construction. Very imprecise results of clustering GIST descriptors in reduced dimensionality require later verification, and extension to neighboring graph nodes. This approach was used in building a system for large, unorganized data sets, however GPS coordinates were used for faster processing.

The problem of similarity search in large image databases is approached in [12]. This work is mostly inspired by results from text search engines ([13], [14]). First a codebook is created by clustering SIFT features. Each image in the database may be described with a set of indices to the cluster centers. The retrieval algorithm searches for nearest cluster centers to the query image features. The most similar images are those containing features associated with cluster centers. The method works with static databases.

6.3 Compact Descriptor and Matching

The key observation is that an image sequence showing the same 3D scene has many common image features visible in different photographs of the sequence. Due to the robustness of SIFT, different viewing angles and distances to the scene have a small influence on SIFT feature descriptors. This fact can be immediately exploited, by using features occurring multiple times in different images of a sequence, just once.

In the following subsections, we describe the similarity descriptor, the method of descriptor matching, and the search algorithm for finding individual matching images among two different image subsets.

6.3.1 Computation of Compact Descriptor

In order to build a descriptor for an image sequence, first, we extract SIFT features from all images. It is possible to limit the number of features by adjusting the contrast threshold. This step is optional, however for plain similarity measurement, it may be worth to limit the input data size for later processing, as it can save computation time for longer sequences.

All features from all images or key frames in a sequence are collected, and then clustered with standard k-means clustering. With this approach, we exploit the fact of having very similar features in many images belonging to the same sequence.

The *compact descriptor* **C** is a set of cluster centers $C_j \in \mathbb{R}^{128}$ of all features from the sequence:

$$\mathbf{C} = \{C_j\}_{j=1}^{k} \tag{6.1}$$

The size of the compact descriptor is $k * 128 *$ sizeof(float), what in most of our tests gives 2048 kilobytes for 4096 cluster centers.

6.3.2 Computation of Similarity Measure

The similarity measure computation can be reduced to standard SIFT feature matching algorithm, as presented in the original publication [2]. We can do this, because a cluster center can be considered as a single SIFT feature, that is an average vector of all features in an image subset.

In the original paper, matching of two sets of features is done as follows: for each feature in the first set, search for the nearest and the second nearest neighbor in the second set. If the ratio of the first nearest neighbor to the distance to the second nearest neighbor is less than a threshold t, then the match is accepted. The choice of $t = 0.8$ was able to eliminate 90 percent of false matches, while discarding less than

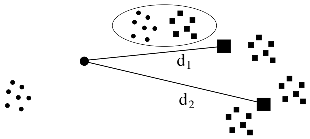


Fig. 6.2 Example configuration of real feature clusters and k-means centers for two image sequences, where the original condition for SIFT feature matching (involving the ratio of d_1 and d_2) may fail.

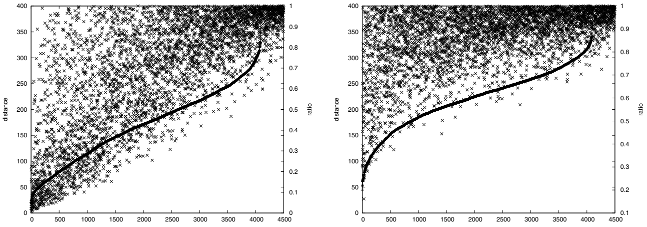


Fig. 6.3 Plots of distances and ratios for a matching sequence (left) and non-matching sequence (right).

5 percent of correct matches. Directly from the original algorithm, we can deliver the following similarity measure between two sequences S_1 and S_2:

$$\sigma(S_a, S_b) = \#\{c \in \mathbf{C}_a | \frac{dist(c, NN_1(c, \mathbf{C}_b))}{dist(c, NN_2(c, \mathbf{C}_b))} < t\} \tag{6.2}$$

what is just a number of matches between cluster centers of two descriptors.

This simple algorithm has proven to be robust in the case of SIFT image feature descriptors, however in case of matching k-means centers, the following problem, illustrated in Figure 6.2, may occur. Due to an insufficient amount of centers used in k-means, one center can represent more than one real cluster of features. A natural consequence of this fact is, that the k-means center is shifted to a location influenced by other centers of gravity of real feature clusters. When matching with the k-means centers from another sequence, the ratio of the first nearest neighbor to the second one may exceed the threshold, with the consequence of being rejected, despite the fact, that the real centers of gravity for clusters of features should fulfill the acceptance condition. In this case, for long sequences with relatively small descriptors, it may be necessary to relax the original condition in order not to throw away close clusters that may contain matching features. We have analyzed plots of the distance between cluster centers together with associated ratios in case of matching and non-matching sequences. It can be clearly seen, that even for close distances, some of the matches have high ratios. We have obtained a good similarity response with the threshold of 0.5 for the ratio of squared distances to the nearest and the second nearest cluster center. Plots for exemplary matching and non-matching sequences, are shown in Figure 6.3.

6.3.3 Computation of Image Occurrence Statistics

As the descriptor itself is sufficient for computing the similarity measure, we are also interested in identifying individual frames that match between image sequences. This is essential for the following purposes:

- Adding single frames to image sequences for obtaining camera calibration parameters, with respect to an already calibrated sequence.
- Recognizing 3D points corresponding to 2D feature points for computing transformations between reconstructed 3D sub-models.
- Adding additional constrains for 3D model alignment and loop closure.

In order to efficiently select pairs of matching frames from different subsets of images, we extend each cluster center C_j of the compact descriptor with a set D_j:

$$D_j = \{I_i\}_{i=1}^{k} \tag{6.3}$$

where I_i is the image number that fulfills the following condition:

$$\exists_{f \in E_j} f \in F_i \tag{6.4}$$

where E_j is a set of image features associated with C_j, and F_i is a set of features associated with image I_i.

In practice, the computation of sets D_j is done by gathering features associated with center C_j and constructing the set of images, from where the features came from. This can be done in the last iteration of the k-means algorithm.

6.3.4 Computation of Matching Frames

Algorithm 1. Computation of Matching Frames

```
function ComputeImageMatches($M, S^1, D^1, S^2, D^2$, cnt)
    $n \leftarrow$ number of images in $S^1$
    $m \leftarrow$ number of images in $S^2$
    allocate array $A$ of size $n \times m$, set to zeros
    for $i = 1 \rightarrow |M|$ do
        $(c^1, c^2, d, r) \leftarrow$ $i$-th 4-tuple of $M$
        for all $d_1 \in D^1_{c^1}$ do
            for all $d_2 \in D^2_{c^2}$ do
                $p_{ok} \leftarrow P(\text{match}(d_1, d_2)|d, r)$
                $p_{nok} \leftarrow P(\text{no-match}(d_1, d_2)|d, r)$
                if $p_{ok} > p_{nok}$ then
                    $A[d_1][d_2] \leftarrow A[d_1][d_2] + 1$
                end if
            end for
        end for
    end for
    return $\{(u^1_i, u^2_i)|A[u^1_i][u^2_i]\text{is } i\text{-th largest element of } A\}_{i=1}^{\text{cnt}}$
end function
```

Matching frames are computed from the image statistics described in the previous subsection. During the similarity computation, for each k-means center c^1_i in the first sequence, we have computed the nearest (c^2_i) and second nearest corresponding cluster center from another image sequence, obtaining set of matches $M = \{(c^1_i, c^2_i, d_i, r_i)\}_{i=1}^{u}$, where d_i is the distance to the nearest center, and r_i is the ratio of squared distances to the nearest and the second nearest cluster center.

The computation of matching images between sequences S^1 and S^2 with corresponding descriptors D^1, D^2, whose elements are defined in Equation 6.3, is shown in algorithm 1.

The idea behind this algorithm is as follows. When we have two matching cluster centers from different image sequences, there is a probability that the images associated with the first center are matching to the images associated with the second center, because they may have at least one common feature. Therefore, we consider all possible image pairs from the two matching centers, and apply a voting mechanism using the array A. Each pair increases an entry in this array depending on probability calculations detailed below. When we consider all matching cluster centers, the voting scheme will cause the most probable matching image pairs to emerge.

The probability function P, describes how probable is that two images from corresponding cluster centers can be matched together. According to the Bayesian theory, this function can be expressed as

$$P(\text{match}(I_1,I_2)|d,r) = \frac{p(d,r|\text{match}(I_1,I_2)) \cdot P(\text{match}(I_1,I_2))}{p(d,r)} \tag{6.5}$$

where the evidence $p(d,r)$ is just a scaling factor, that can be omitted in the Bayesian decision rule. As the prior probability $P(\text{match}(d_1,d_2))$ is unknown, at this point we need to assume it equal with $P(\text{no-match}(d_1,d_2))$. The simplified decision rule now depends only on the likelihoods $p(d,r|\omega)$, where ω is the matching and non-matching class. In order to estimate those two density functions $p(d,r|\omega)$, the idea that immediately comes to a mind is to use a standard technique from statistics - kernel density estimation. It would be required to gather a lot of data about distances and ratios of matching and non-matching pairs, and estimate the probability density directly from this data. However, we used instead an extremely simplifying but effective approximation:

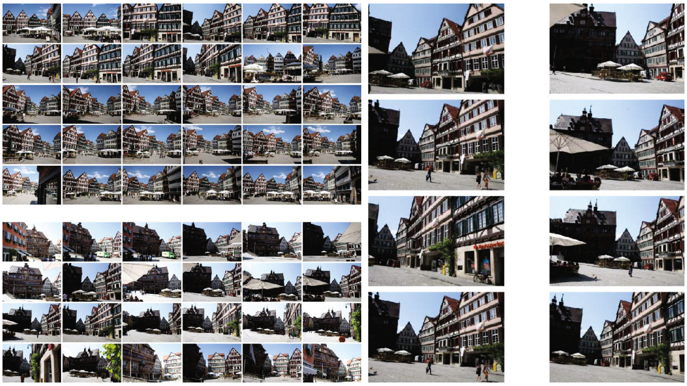
$$p(d,r|\text{match}(I_1,I_2)) = \begin{cases} 2, & r \in (0,\frac{1}{2}] \\ 0, & r \notin (0,\frac{1}{2}] \end{cases} \tag{6.6}$$

According to the above formula, the decision about considering a possible image pair as matching, in practice turns down to checking, if the ratio associated with matching k-means centers falls into the interval $(0,\frac{1}{2}]$. The choice of the interval is directly connected to the sequence matching threshold for k-means centers, as described in Chapter 6.3.2. This simple approximation would not be sufficient for a correct decision when a single image pair associated with two matching clusters is taken into consideration, however when used in the voting scheme, it provides stable results.

We remove the most frequently occurring image pair from the array A. This helps to filter out false positives from the query response in the context of image search databases, as it is suggested in [12] and supported by our experiments. We do not check the geometric configuration of features within the images. This simplification can limit the accuracy of the algorithm, however any false positives are detected in the later stage of feature matching between proposed matching image pairs.

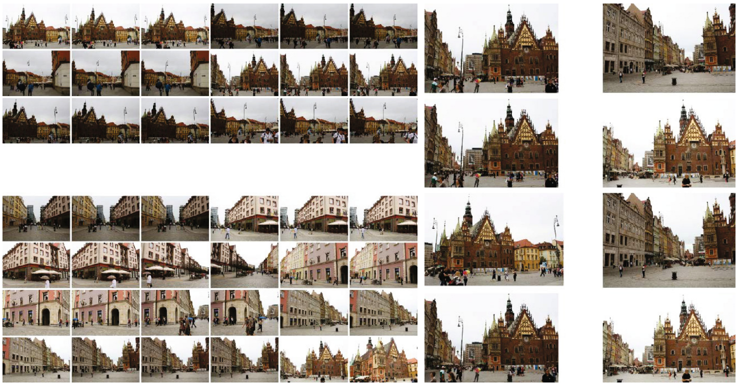
6.4 Results

We have tested our compact descriptor on many datasets. Example datasets are shown in Figures 6.4, 6.5 and 6.6 (a). In the same Figures (b), we show examples of matching pairs. Pairs containing repeating images from two sequences have been omitted for space saving reasons, and the most highly ranked pairs are shown. The resolution of images used in experiments is 10 mega pixels, From each image, we



(a) (b)

Fig. 6.4 Image matching results for Tue1 dataset. Two sequences are shown on (a), and matching pairs on (b).



(a) (b)

Fig. 6.5 Image matching results for Wro3 dataset. Two sequences are shown on (a), and matching pairs on (b).

use 1000 features with the highest contrast. Features are gathered and clustered to the descriptor size of 4096 k-means centers.

The time required for image sequence matching is comparable to SIFT feature matching between two images, and on a single core of an Intel Core Quad Q9300 running at 2.5 GHz, is approximately 5 seconds. No optimization or parallelization of the image sequence matching code has been performed yet. The creation of compact descriptors can be done nearly as fast as SIFT feature extraction, for tested sequences of 45 images, it takes approximately 223 seconds (4 threads). In our experiments, we use the maximum of 32 k-means iterations. The descriptor creation time is much longer than the matching time, however it is created just once, and it is reused many times in the course of large scale reconstruction.

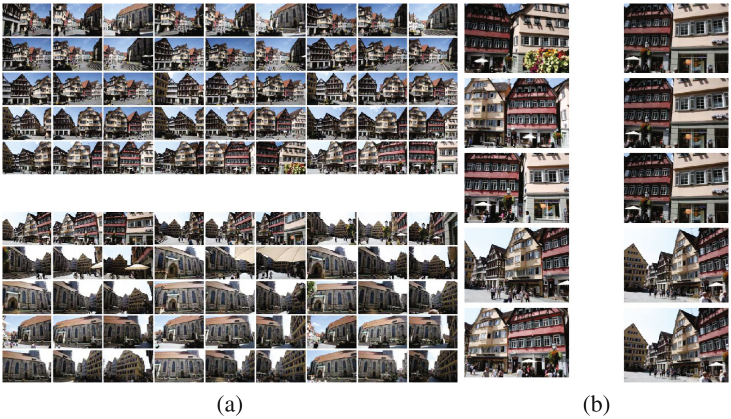


Fig. 6.6 Image matching results for Tue2 dataset consisting of two 45 image sequences. Two sequences are shown on (a), and example matching pairs on (b).

False positive matches are ranked low, and they have never appeared in the first 16 matching pairs used for assembling a final 3D model. However, if any false positive matches appeared, they would be discarded in the later stage of geometric verification of the reconstruction algorithm.

The algorithm for similarity descriptor matching has been proven in creating correct similarity graphs between image sequences used to merge reconstructed 3D sub-models. One of the results can be seen in Figure 6.1.

6.5 Conclusions and Future Work

We have presented a novel approach for video sequence matching using compact descriptors. The speed of our algorithm in matching two video sequences can be compared to the speed of the standard SIFT feature algorithm for two images. We have successfully used the described descriptors in context of large scale reconstruction.

I future work, we would like to address the following issues:

- Additionally to the selection of individual matching images, we will work on extending the algorithm to output individual features that match in the selected images. This would require checking individual features contained in corresponding k-means clusters.
- We will work on an improved and still fast probability density function and probability integration instead of the simple voting scheme. However the estimate in Equation 6.6 provides stable results, and does practically not contribute to total matching time.

References

1. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2006, vol. 1, pp. 519–528. IEEE Computer Society, Washington, DC (2006)
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
3. Hampapur, A., Hyun, K., Bolle, R.M.: Comparison of sequence matching techniques for video copy detection. In: Storage and Retrieval for Media Databases, pp. 194–201 (2002)
4. Kim, Y.-T., Chua, T.-S.: Retrieval of news video using video sequence matching. In: Proceedings of the 11th International Multimedia Modelling Conference, MMM 2005, pp. 68–75. IEEE Computer Society, Washington, DC (2005)
5. Kim, S.H., Park, R.-H.: An efficient algorithm for video sequence matching using the modified hausdorff distance and the directed divergence. IEEE Trans. Circuits Syst. Video Techn. 12(7), 592–596 (2002)
6. Chen, L., Stentiford, F.W.M.: Video sequence matching based on temporal ordinal measurement. Pattern Recogn. Lett. 29(13), 1824–1831 (2008)
7. Yeh, M.-C., Cheng, K.-T.: Video copy detection by fast sequence matching. In: Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR 2009, pp. 45:1–45:7. ACM, New York (2009)
8. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Comput. Vis. Image Underst. 110(3), 346–359 (2008)
9. Raginsky, M., Lazebnik, S.: Locality-sensitive binary codes from shift-invariant kernels. In: NIPS, pp. 1509–1517 (2009)
10. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
11. Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S., Pollefeys, M.: Building Rome on a Cloudless Day. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 368–381. Springer, Heidelberg (2010)
12. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: IEEE International Conference on Computer Vision (2007)
13. Risvik, K.M., Aasheim, Y., Lidal, M.: Multi-tier architecture for web search engines. Web Congress, Latin American, 132 (2003)
14. Barroso, L.A., Dean, J., Hölzle, U.: Web search for a planet: The google cluster architecture. IEEE Micro. 23(2), 22–28 (2003)

Chapter 7
Headlines Usefulness for Content-Based Indexing of TV Sports News

Kazimierz Choroś

Abstract. In the classical indexing process of text documents keywords are derived mainly from the title, chapter titles, figure legends, table captions, and other special part of a text. The same strategy seems to be adequate also for a video indexing. The content analysis is more effective when the structure of a video is taking into account. A digital video similarly to text document is also hierarchically structured into a strict hierarchy. It is composed of different structural units such as: acts, episodes (sequences), scenes, camera shots and finally, single frames. The sequence of scenes in a video is usually organized in a standard way typical for a given category of a video. Particularly TV shows are edited respecting standard rules. The chapter presents the results of analyses of the structure of TV sports news and of the usefulness of sport headlines for content-based video indexing. The sport headlines and the video editing schemes recognized for a given video type may significantly help to reduce the number of frames analyzed during content-based indexing process.

7.1 Introduction

New technologies and new methods of indexing applied in visual retrieval systems allow the storage and retrieval of a very huge amount of digital video data. Video data have become publicly and relatively easy available. The methods used in

Kazimierz Choroś
Institute of Informatics, Wrocław University of Technology
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: kazimierz.choros@pwr.wroc.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 65–76.
springerlink.com

textual retrieval systems cannot be adapted in visual retrieval systems. Visual data are composed of and cannot be indexed in traditional ways, i.e. by the index terms derived directly form the indexed data. We need the special method for content analyses of visual data. Manual indexing is unfeasible for large video collections. Many approaches, frameworks, methods, algorithms of automatic indexing and retrieval of visual information have been proposed and are still being developed. The content-based automatic indexing and retrieval of video data are very complex processes, because content is very subjective to be characterized easily and completely. Content-based indexing of videos has become a research topic of increasing importance. New methods for organizing, indexing, browsing, and retrieval of videos in large video archives use high-level semantic features. Therefore, we are looking for effective tools to identify the video segments with a specific content, for example news on weather, sports, science, finances, technology, world travel, national economy, or entertainment news.

The automatic detection and categorization of sport events in TV sports news is one of the main area of content-based video indexing. The analyses of sports news are time-consuming processes because in most cases all frames of videos are individually processed. It would be desirable to reduce the analyses to some parts of videos, similarly to text indexing when we limit the analysis of the text to the most informative parts of the text document lie title and abstract. In general, text documents have formal structure. It enables us to optimize the processing of the text. The question arises if the detection of the video structure may also lead to the more effective solution in content-based video indexing. The answer seems to be obvious. So, which structural elements are the most useful for video indexing?

The recognition of video structure requires the application of video segmentation procedures and on the other hands the detection of relations between shots and scenes in the video.

The chapter is organized as follows. The next section describes the main related works in the area of automatic categorization of shots in TV sports news videos. Some recent related research works are cited and their main ideas are presented. The section 3 discusses analogies of text and video structures. The juxtaposition is also presented of two indexing processes, i.e. of text and video indexing based on the content analysis of their structure units. The comparison of a structure of a book, journal paper, and TV sports news is also presented in this section. The section 4 presents the Automatic Video Indexer AVI which is a research project investigating tools and techniques of automatic video indexing for retrieval systems. The section 5 presents a standard structure of TV sports news. Then the most informative parts of a text document and of sports news are compared. In the sixth section the usefulness of headlines preview for video indexing is considered and analysed. The final conclusions and the future research work areas are discussed in the last section of this chapter.

7.2 Related Works

Last years many investigations in automatic recognition of a content of a video clip [1-3] have been carried out, many of proposed methods have been tested on sport videos. An automatic summarization of TV sport videos has become a popular application because sport videos are extremely popular in all video databases and Web archives. Because of a huge commercial appeal sports videos became a dominant application area for video automatic indexing and retrieval.

Segmentation of a video leads to the identification of the standard basic video units, such as shots and scenes [4, 5]. These formal video structure units are very useful in video editing process. But the identification of logic elements, sections, parts of information in a video can optimize the content-based indexing process [6, 7]. Also usefulness of news headlines have been examined in indexing process [8].

News video can be divided into six sections such as follows: opening animation, anchorperson greeting, headlines preview, news stories, weather forecast and the closing [9]. Experimental results have shown that the segmentation of the news videos into the parts mentioned above can the user to quickly browse or retrieve segments of interest from the news videos.

Many experiments have been also performed on the categorization of sports events detected in sports news [4, 10-13] and many approaches and schemes have been developed. For example a unified framework for semantic shot classification in sports videos has been defined in [14]. The proposed scheme makes use of domain knowledge of specific sport to perform a top-down video shot classification, including identification of video shots classes for each sport. The method has been tested over 3 types of sports videos: tennis, basketball, and soccer.

Other experiments have been carried out for example with soccer [15], baseball videos [16], with tennis videos [17, 18], as well as with other sports. There are also many promising experiments in which the specific features of sport courts (for example lines) [19, 20] or sports equipments [20] (for example ball, bicycle, or tennis racket) are used to classify sport events in videos.

7.3 Book Structure vs. TV Sports News Structure

Text is autodescriptive. The text in any natural language is composed of words, sentences, paragraphs, and chapters. The process of indexing is usually limited to the identification of words or expressions in the text. These words or expressions are used as index terms in retrieval systems. Let's notice that the text has also structural features. These are for example: a language of the text, its length measured in characters, but also in words, lines, or pages.

A digital video is also hierarchically structured into a strict hierarchy [21, 22]. It is composed of the following structural units: acts, episodes (sequences), scenes, camera shots and finally, single frames. A shot, a basic unit is usually defined as a continuous video acquisition with the same camera, so, it is a sequence of interrelated consecutive frames recorded contiguously and representing a continuous action in time or space. Depending on the editing style shots in a given scene are content related but they can be temporally separated and/or even spatially disconnected.

Table 7.1 Analogies in the structures of a text and of a video [22]

Text	Video
character	frame
word	shot
sentence	scene
paragraph	episode
chapter	act
book	movie

The strategy of indexing process can be undertaken similar to that of the text indexing, form the simplest textual units to the most advanced parts. Table 7.1 compares the basic textual units with video structural elements and Table 2 (presented already in [23]) compares two indexing processes based on the content analysis of their structure units.

The great analogies between the structural elements of a text and the structural elements of a video clip can be observed. The video can be seen as a visual representation of the book, something like a novel and a movie using the screenplay written on the basis of this novel.

Table 7.2 Comparison of processes of text indexing and of video indexing [22]

Text Indexing	Video Indexing
character decoding: recognition of individual characters, elimination of punctuation symbols	frame decoding: frame analysis, calculation of frames characteristics (histograms etc.)
word selection in a text: morphological analysis, elimination of words using stop-lists, word normalisation, identification of descriptors from a thesaurus, identification of relation between words, calculation of word frequency	temporal segmentation: shot detection, calculation of shot length, shot filtering – elimination of shots too short, content-based categorisation of shots, detection of objects, faces, lines, words etc. in a shot
morpho-syntactic analysis of sentences: identification of multi-word (compound) index terms, syntactical patterns, multi-word expressions, noun phrases	scene detection: scene filtering, shot clustering, pattern analysis of scenes
semantic analysis of a paragraph: semantic and contextual analysis of sentences	content analysis of episodes: scene clustering
content analysis of chapters and of the whole text (book)	content analysis of acts and of the whole movie

The comparison of various components of a book and of structural elements of a video (Table 7.3) leads to the conclusion that these different media have analogous structures, so, the indexing strategies should be similar in some extent.

Table 7.3 Comparison of the structures of a book and of TV sports news

Book structure	Journal paper structure	TV sports news structure
front cover: title and/or author, usually with possibly an appropriate illustration	title and author: paper title, author name	3D intro animation 3D computer graphics production logo, video title, title sequence
edition notice or copyright page: copyright notice, legal notices, publication information, printing history, cataloguing information, ISBN – International Standard Book Number	author affiliation	
front matter: dedication, contents list or table of contents, foreword or preface, acknowledgments, introduction, prologue	abstract	headlines: sports news headlines
body matter: volumes, parts, chapters and sections	paper text: paper sections	sport events: best sport highlights of recent top sports events
conclusions	conclusions: conclusions, main results, further research	anchorman announcements: anchorman announcing most important forthcoming sport events and saying goodbye
back matter: bibliography, list of figures and list of tables (they may be included in the front matter), possibly appendix, glossary, index, colophon, possibly errata	bibliography	3D final animation: usually the same as 3D intro animation but with a superimposed text (editor name and producer name)

It should be noticed that a book is traditionally a sequential text, in contrast to hypertext which overcomes the traditional linear constraints of written text. Traditional video has also a sequential nature, although semantically the scenes of the same episode in a given video are not necessarily placed one after another. But such a situation is also natural for a written text in a novel. It usually happens that two or several episodes are presented in alternation.

7.4 AVI – Automatic Video Indexer

The AVI – Automatic Video Indexer [24] is a research project investigating tools and techniques of automatic video indexing for retrieval systems. The main goal of the project is to develop efficient techniques of content-based video retrieval. All tests will be performed in the AVI Indexer.

Two main processes already implemented in the Automatic Video Indexer are: the Automatic Shot Detector ASD responsible for temporal segmentation and shot categorisation and the Automatic Scene Analyser ASA responsible for shot clustering, scene detection, and content analysis of scenes. The modules being developed will identify players, playing fields, and sport equipments. The goal is to extract the most interesting highlights, which facilitate browsing and retrieval of sports video.

The first step of content-based video indexing in the AVI Indexer is a temporal segmentation leading to the segmentation of a movie into small units called video shots. There many methods proposed for temporal segmentation and in general the process is well managed. In the next step the key frame are extracted. The key frames should be the best for depicting the content of corresponding shot or scene. Then in the third step the content of the shots detected during the temporal segmentation is analysed. The content can be identified using different approaches: comparison of frames with image patterns, line detection in playing fields, detection of superimposed text, face detection of players, detection of sport objects, detection of player and audience emotions. Another problem is shot clustering, and in consequence semantic scene segmentation.

7.5 Structure of TV Sports News

It was observed that TV sports news program has a specific structure. The analyses of TV sports newscast broadcasted in the first national Polish TV channel (TVP1) show that it has its individual standard editing structure. It is composed of several highlights introduced and commented by anchorperson, and often accompanying by numerical results presented in tables. Figure 7.1 presents the most typical structure of TV sports news.

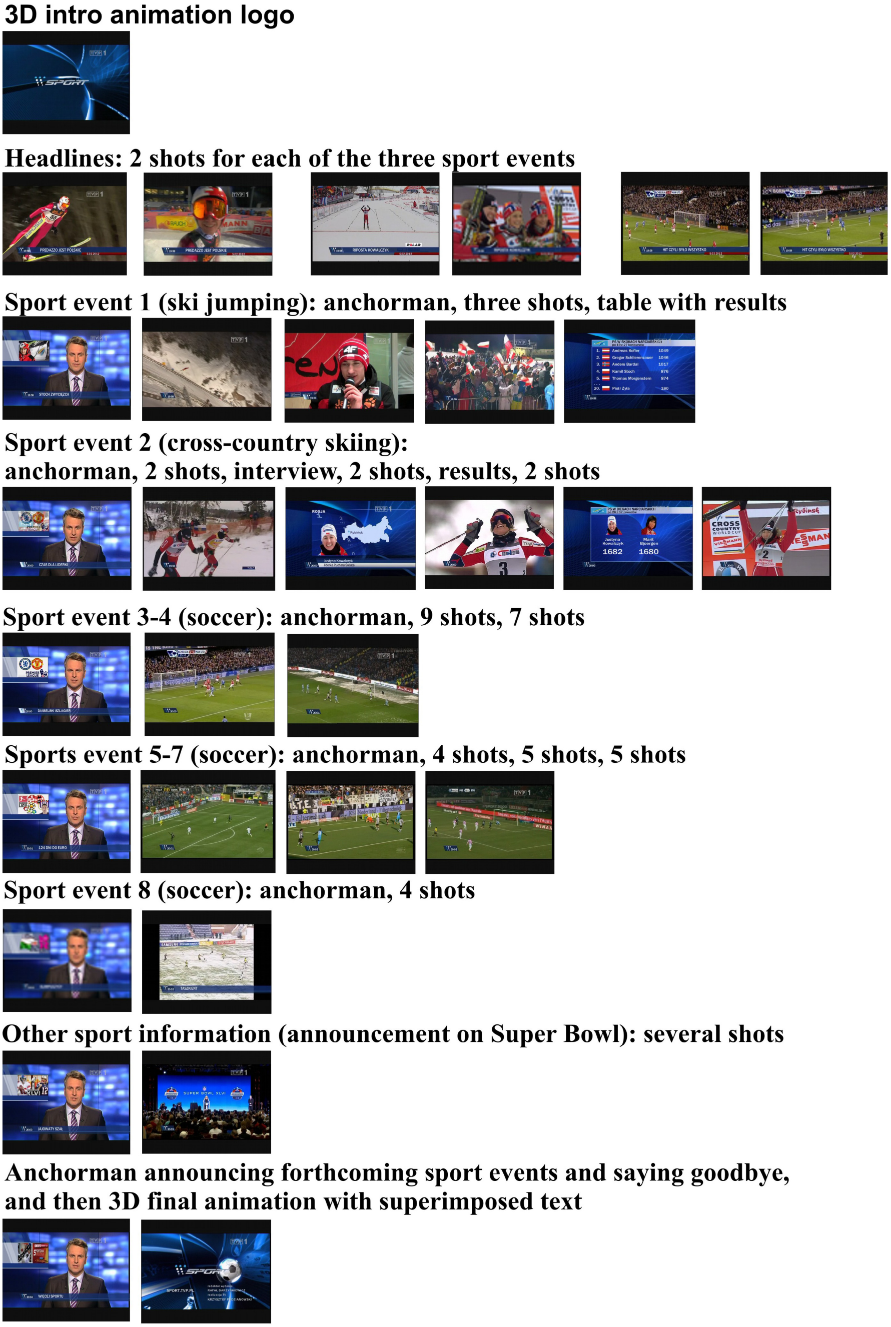

3D intro animation logo
Headlines: 2 shots for each of the three sport events
Sport event 1 (ski jumping): anchorman, three shots, table with results
Sport event 2 (cross-country skiing):
anchorman, 2 shots, interview, 2 shots, results, 2 shots
Sport event 3-4 (soccer): anchorman, 9 shots, 7 shots
Sports event 5-7 (soccer): anchorman, 4 shots, 5 shots, 5 shots
Sport event 8 (soccer): anchorman, 4 shots
Other sport information (announcement on Super Bowl): several shots
Anchorman announcing forthcoming sport events and saying goodbye,
and then 3D final animation with superimposed text

Fig. 7.1 The structure of TV sports news on 5 Feb., 2012 – typical structure of TV sports news.

The knowledge on video structure can suggest which parts of video are probably the most informative, so, frames from which parts of a video should be selected for content analyses and for the content-based indexing process. The comparison of the most informative parts of a text and their analogous counterparts in a video (Table 7.4) may be a strategic indication for such decision.

Table 7.4 Comparison of (analogies in) the most informative structural parts of a text and of TV sports news video

Main parts of text document	Main parts of video
full text	whole video
title	3D intro but without content information
abstract	headlines
table of contents	anchorman sometimes announces what sport events will be presented but this fact is not reflected in video
introduction	anchorman self-presentation
chapter title	first frame of a scene (or first frame of the first scene of a given sport event)
table caption	sporting highlight
figure legend	sporting highlight
highlighted sentence or paragraph	sporting highlight
highlighted word	first frame of a shot
conclusions	final studio scene

7.6 Headlines in TV Sports News

The most informative parts of a video has not exact adequate element in the textual document structure. Nevertheless, similarly to text indexing strategy it is reasonable to define the most informative parts in the video structure. Headlines seem to be the best part of TV sports news for automatic content-based indexing. Tables 7.5 and 7.6 present the main characteristics of headlines and their usefulness for the indexing of TV sports news videos.

The first observation is that the number of sport events reported does not directly depend on the duration of TV sports news broadcast.

The content-based analysis basing only on headlines can ensure 53% of the recall of the detection of sport events presented in TV sports news (Table 7.6). The analyses of frames only in headlines do not permit to index all sport events reported in the TV sport news, although, the main events would be indexed after processing of only about 6 shots. Assuming that in most cases one frame is sufficient to categorize sport shot the analysis of only 6 frames is sufficient to detect more than half of sport disciplines in TV sports news.

Table 7.5 Sports events in headlines of TV sports news

	Date	Duration [min:sec]	Number of shots in headlines	Number of sport events in headlines	Number of sport disciplines in headlines	Number of sport events in a TV sports news broadcast	Number of sport disciplines in a TV sports news broadcast
1.	Feb. 01, 12	10:25	8	3	3	10	3
2.	Feb. 02, 12	8:15	7	3	3	8	4
3.	Feb. 03, 12	8:15	7	3	3	9	6
4.	Feb. 04, 12	8:15	7	3	3	9	6
5.	Feb. 05, 12	6:25	6	3	3	9	4
6.	Feb. 06, 12	6:40	6	3	3	7	4
7.	Feb. 07, 12	6:15	7	4	3	7	6
8.	Feb. 08, 12	9:30	6	3	3	9	4
9.	Feb. 09, 12	9:20	6	3	1	12	4
10.	Feb. 10, 12	9:40	4	3	3	8	6
11.	Feb. 11, 12	9:25	8	3	3	11	8
12.	Feb. 12, 12	9:00	5	3	3	16	11
13.	Feb. 13, 12	7:50	7	3	2	11	3
14.	Feb. 14, 12	8:50	7	4	3	10	4
15.	Feb. 15, 12	8:30	7	3	2	8	4
16.	Feb. 16, 12	8:35	8	6	3	13	5
17.	Feb. 17, 12	9:30	5	3	3	9	6
18.	Feb. 18, 12	3:50	6	3	3	5	5
19.	Feb. 19, 12	6:45	6	3	3	8	6
20.	Feb. 20, 12	6:40	7	6	1	12	4
21.	Feb. 21, 12	4:25	6	3	2	4	2
22.	Feb. 22, 12	5:30	7	3	2	5	3
23.	Feb. 23, 12	9:10	6	4	2	9	4
24.	Feb. 24, 12	6:50	6	3	3	6	4
25.	Feb. 25, 12	4:15	7	3	3	6	5
26.	Feb. 26, 12	7:25	7	3	2	13	10
27.	Feb. 27, 12	6:20	4	2	2	8	3
28.	Feb. 28, 12	3:50	3	1	1	5	1
29.	Feb. 29, 12	6:45	7	1	1	2	1
	Σ	216:25	183	91	72	249	136

Table 7.6 Statistical analysis of headlines in TV sports news of the First national Polish TV channel (TVP1) broadcasted in February 2012 (29 broadcasts)

Characteristics of TV sports news in Polish First national channel (TVP1) in February 2012	
Total number of broadcasts of TV sports news	29
Average duration of TV sports news	7:28
Average number of shots in headlines	6.31
Average number of sport events in headlines	3.14
Average number of sport disciplines in headlines	2,48
Average number of shots in headlines announcing one sport event	2.01
Average number of all sport events presented in one broadcast of TV sports news	8.59
Average number of different sport disciplines in one broadcast of TV sports news	4.69
Average number of sport events not-announced in headlines	5.45
Recall of sport events in headlines	37 %
Recall of sport disciplines in headlines	53 %

7.7 Final Conclusions and Further Studies

Sport news media are one of the most requested subjects in the Web. Many people are sports fanatic, they read almost everything about sports, sports fans like to see many times the best highlights of their favourite sports games. TV sports news videos are relatively short. The standard duration is several minutes. Nevertheless, it would be desirable to limit the time-consuming analyses of contents to the most informative parts of a broadcast. For this purpose we need to recognize the structure of TV sports news.

The great similarity of text and video structure could suggest that we may follow the same strategies as in the case of textual retrieval. In a text document the main parts, the most informative, and the most useful for indexing and retrieval are title and abstract. For many years information retrieval systems indexed documents only on the basis of text analyses of title and abstract.

Unfortunately, the analyses of TV sport news broadcasted in the first Polish national TV channel have shown that videos such as TV sports news do not have informative title. The title of TV sports news, i.e. animated logo, informs only on the genre of TV broadcast but says nothing on its content. Headlines in TV sports news which seem to be similar to an abstract of a textual document ensure the detection of only just over half of sport disciplines in a broadcast.

However, the knowledge of the structure of TV sports news can be useful and can significantly reduce the number of analyzed frames. The indexing process can be limited to the video headlines presented at the beginning of TV sports news and to only one frame from the shot following the studio shot, i.e. the long shot with anchorman announcing the next sport event reported in the news.

In further research and experiments new computing techniques will be developed for the Automatic Video Indexer. Its functionality will be extended by introducing an automatic extraction of video features and objects like faces, lines, texts, etc., as well as its application will be extended to other kinds of TV shows.

References

1. Geetha, P., Narayanan, V.: A Survey of content-based video retrieval. Journal of Computer Science 4(6), 474–486 (2008)
2. Money, A.G., Agius, H.: Video summarisation: a conceptual framework and survey of the state of the art. Journal of Visual Communication and Image Representation 19, 121–143 (2008)
3. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A survey on visual content-based video indexing and retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 41(6), 797–819 (2011)
4. Choroś, K.: Video Shot Selection and Content-Based Scene Detection for Automatic Classification of TV Sports News. In: Tkacz, E., Kapczynski, A. (eds.) Internet – Technical Development and Applications. AISC, vol. 64, pp. 73–80. Springer, Heidelberg (2009)
5. Tapu, R., Zaharia, T.: High Level Video Temporal Segmentation. In: Bebis, G. (ed.) ISVC 2011, Part I. LNCS, vol. 6938, pp. 224–235. Springer, Heidelberg (2011)
6. Han, H., Kim, J.: An useful method for scene categorization from new video using visual features. In: Third World Congress on Nature and Biologically Inspired Computing (NaBIC), pp. 480–484 (2011)
7. Rao, K.S., Pachpande, K., Vempada, R.R., Maity, S.: Segmentation of TV broadcast news using speaker specific information. In: Proceedings of the National Conference on Communications (NCC), pp. 1–5 (2012)
8. Jia, Y., Chen, Z., Yu, S.: Reader emotion classification of news headlines. In: Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 1–6 (2009)
9. Ko, C.-C., Xie, W.-M.: News video segmentation and categorization techniques for content-demand browsing. In: Proceedings of the Congress on Image and Signal Processing (CISP 2008), vol. 2, pp. 530–534 (2008)
10. Yang, Y., Lin, S.-X., Zhang, Y.-D., Tang, S.: Statistical Framework for Shot Segmentation and Classification in Sports Video. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 106–115. Springer, Heidelberg (2007)
11. Li, L., Zhang, N., Duan, L., Huang, Q., Du, J., Guan, L.: Automatic sports genre categorization and view-type classification over large-scale dataset. In: Proceedings of the seventeen ACM International Conference on Multimedia, pp. 653–656 (2009)
12. Choroś, K., Pawlaczyk, P.: Content-Based Scene Detection and Analysis Method for Automatic Classification of TV Sports News. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS(LNAI), vol. 6086, pp. 120–129. Springer, Heidelberg (2010)
13. Zhang, N., Guan, L.: An efficient framework on large-scale video genre classification. In: IEEE International Workshop on Multimedia Signal Processing (MMSP), pp. 481–486 (2010)

14. Ling-Yu, D., Min, X., Qi, T., Chang-Sheng, X., Jin, J.S.: A unified framework for semantic shot classification in sports video. IEEE Transactions on Multimedia, 1066–1083 (2005)
15. Ishida, K., Tanaka, M.: Identification of the part of soccer court from video signal by neural networks. In: Proceedings of the Int. Conf. on Control, Automation and Systems 2008, pp. 2563–2568 (2008)
16. Lien, C.-C., Chiang, C.-L., Lee, C.-H.: Scene-based event detection for baseball videos. Journal of Visual Communication and Image Representation 18, 1–14 (2007)
17. Huang, Y., Choiu, C., Sandnes, F.E.: An intelligent strategy for the automatic detection of highlights in tennis video recordings. Expert Systems with Applications 36(6), 9907–9918 (2009)
18. Huang, Y., Choiu, C., Sandnes, F.E.: An intelligent strategy for the automatic detection of highlights in tennis video recordings. Expert Systems with Applications 36, 9907–9918 (2009)
19. Cai, Z.Q., Tai, J.: Line detection in soccer video. In: Proceedings of the Fifth International Conference on Information, Communications and Signal Processing, pp. 538–541 (2005)
20. Li, Y., Liu, G., Qian, X.: Ball and field line detection for placed kick refinement. In: Proc. Global Congress on Intelligent Systems (GCIS), vol. 4, pp. 404–407 (2009)
21. Geng, Y., Xu, D., Feng, S.: Hierarchical Video Summarization Based on Video Structure and Highlight. In: Yeung, D.-Y., Kwok, J.T., Fred, A., Roli, F., de Ridder, D. (eds.) SSPR 2006 and SPR 2006. LNCS, vol. 4109, pp. 226–234. Springer, Heidelberg (2006)
22. Zhang, S., Zhang, Y., Chen, T., Hall, P.M., Martin, R.: Video structure analysis. Tsinghua Science and Technology 12(6), 714–718 (2007)
23. Choroś, K.: Video structure analysis for content-based indexing and categorisation of TV sports news. Int. J. Intelligent Information and Database Systems 6 (in press, 2012)
24. Choroś, K.: Video Structure Analysis and Content-Based Indexing in the Automatic Video Indexer AVI. In: Nguyen, N.T., Zgrzywa, A., Czyżewski, A., et al. (eds.) Advances in Multimedia and Network Information System Technologies. AISC, vol. 80, pp. 79–90. Springer, Heidelberg (2010)

Chapter 8
Examining Classifiers Applied to Static Hand Gesture Recognition in Novel Sound Mixing System

Michal Lech, Bozena Kostek, and Andrzej Czyżewski

Abstract. The main objective of the chapter is to present the methodology and results of examining various classifiers (Nearest Neighbor-like algorithm with non-nested generalization (NNge), Naive Bayes, C4.5 (J48), Random Tree, Random Forests, Artificial Neural Networks (Multilayer Perceptron), Support Vector Machine (SVM) used for static gesture recognition. A problem of effective gesture recognition is outlined in the context of the system based on a camera and a multimedia projector enabling a user to process sound in audio mixing domain by hand gestures. The image processing method and hand shape parameterization method are described in relation to the specificity of the input and data classifiers. The SVM classifier is considered the optimum choice for the engineered gesture-based sound mixing system.

8.1 Introduction

Reliable gesture recognition in a video stream in most cases demands computationally expensive methods of image processing and object classification [2, 16, 22]. Combining several image processing methods enables to obtain precise localization and shape of a hand in the image. In association with a cascade of algorithms of objects and event classification, theoretically one can obtain high efficacy of gesture recognition. However, using several computationally expensive methods results in decreased efficiency of the whole system when running on standard PC resources. Using such a technology becomes impractical in many applications, since the system is not capable of processing video frames with

Michal Lech · Bozena Kostek · Andrzej Czyzewski
Multimedia Systems Department, Gdansk University of Technology
Narutowicza 11/12, 80-233 Gdansk, Poland
e-mail: mlech@sound.eti.pg.gda.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 77–86.
springerlink.com

sufficient speed. As the result, rapid hand movements associated with dynamic gestures or fast changes of palm shapes associated with static gestures are not fully recorded. This can lead to erroneous interpretation of gestures. Additionally, using such a gesture classification system becomes problematic when considering interaction in which speed of the transitions is essential and performing a gesture affects system continuously and not by steps.

An example of such an application is the gesture-based sound mixing system outlined in the chapter. The process of sound mixing is the art of combining and processing a number of audio signals together to create a "mix". In the case of the engineered system control over sound parameters is provided employing hand gestures. It is of a crucial importance that the system provides instantaneous reaction to the gesture performed, manifesting in intended change to the mixed sound. Therefore, the optimum image processing and gesture classification methods have been chosen with a view to finding a compromise between recognition reliability and system efficiency. In particular, much effort can be put into choosing an optimum classifier as there exist many methods offering various performance in the sense of both computational cost and classification efficacy. Moreover, the classifier performance may change depending on the specificity of the problem to be solved [11].

8.2 System Overview

The components of the engineered system are: a PC or laptop (2 GHz dual core), a webcam, a multimedia projector and a screen. The localization of a user relative to the components is presented in Fig. 8.2.1. The user is localized in a so-called 'sweet spot' (*the focal point of sounds between two speakers*) between the multimedia projector and the projection screen. The projector is hanged under ceiling on such a height and such a distance from the screen that outstretched arms cast shadow on the screen. The camera is placed in front of the user and directed at the screen. The camera lens is set in such a way that the image displayed by the projector is of the greatest size possible when viewed in the frame.

Both dynamic gestures (motion trajectory) and static gestures (palm shapes) are recognized by the system. The method of dynamic gestures recognition has been presented in other works of the authors [14, 15]. Dynamic gestures are closely associated with static gestures. Thus, performing the same motion with a palm shaped differently has various meanings. Moreover, the order in which gestures are performed represents a gesture class. Each recognized gesture is interpreted according to the associated system event, i.e. emulating key or mouse button pressing. The events are received by the engineered sound mixing interface (Fig. 8.2.1) which translates them to MIDI messages being sent to the chosen DAW (*Digital Audio Workstation*) software. Thus, modification of seven audio parameters, namely level, panorama, gain of first order treble shelving filter, dynamic compression's ratio and threshold, and reverb time and mix, is possible.

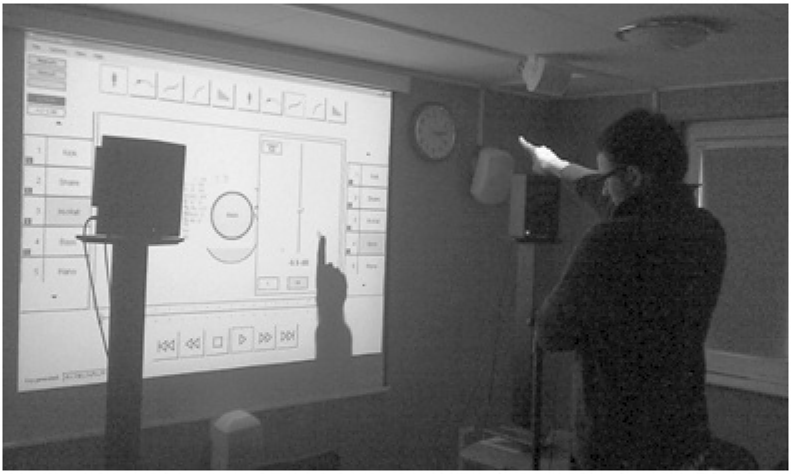


Fig. 8.2.1 Placement of speakers and location of a user during system handling

The image processing method can significantly affect the gesture recognition efficacy as it determines quality of palm shape masks constituting the input to classifiers. Therefore, the method enabling to obtain noise filtered processed video stream input was engineered and presented herein.

The solution bases on subtracting the image extracted from the video stream from the image displayed by the multimedia projector. Gestures are recognized in further processed output. In the first step, the effective area of the image captured from the camera is determined. This area contains only the image displayed by the projector and is determined by the user who points out positions of the image corners in the frame. Based on these positions, the projected image is scaled to ensure identical dimensions with the camera frame. In the next step, the perspective correction is performed. To reduce impact of light conditions and distortions introduced by the camera lens, especially vignetting effect, color calibration is performed. During this process five solid color pictures (red, green, blue, white and black) are displayed. Each camera frame respectively for each background color is subtracted from particular displayed image. Based on the results, tables of discrete constant values used in the later image processing are created. Gesture recognition iteration begins with subtracting processed camera frames from projected images. Subtraction is done in the RGB color space. To each pixel of the output image an appropriate value retrieved from the color calibration tables is added. This value is chosen based on displayed screen particular pixel component intensity. The result is binary thresholded and median filtered. The obtained image is the input to the contours finding method [3, 8, 20] implemented in OpenCV library [6] utilized in the implementation of the system. The method enables to detect hands and eliminates accidental blobs.

8.3 Examining Classifiers Performance

The efficacy of static gesture recognition in vision-based systems highly depends on the type of the shape classification method. This determines the global system efficacy, usability and ease of learning. For this reason, it is essential to choose the best possible classifier. The tested classifiers were: Nearest Neighbor-like algorithm with non-nested generalization (NNge) [7], Naive Bayes, C4.5 (J48) algorithm, Random Tree, Random Forests, Artificial Neural Networks (Multilayer Perceptron) and Support Vector Machines (SVM) from LibSVM library [9]. These particular classifiers have been chosen because of their satisfying results in qualitative assessment for similar problems solving found in literature [4, 5 12, 13, 19, 21, 23, 24]. The implementations of the classifiers originated from the WEKA system [10].

8.3.1 Tests

A group of people who took part in the tests consisted of 18 persons (5 female, 13 males). Each person was asked to perform four predefined static gestures, presented in Fig. 8.3.1, each with three different motion trajectories, i.e. moving a hand from the left to right side alternately, moving a hand up and down alternately, and moving a hand in a gesticulation of circle drawing. Each trajectory was represented by 30 camera frames. For shape parameterization we used the PGH (*Pairwise Geometrical Histograms*) method [6]. This algorithm was chosen particularly because it preserves invariant feature distribution regardless of the palm rotation. It is consistent with the assumption that the rotation angle should not be a factor separating gesture classes and therefore should not affect recognition efficacy. Additionally, the method is insensitive to size of the parameterized shape. Thus, the localization of a user relative to sweet spot does not affect the gesture recognition efficacy.

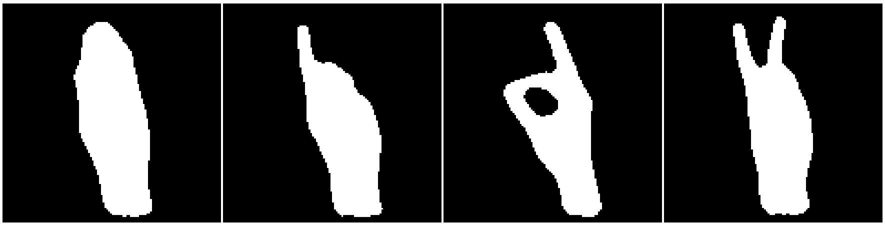


Fig. 8.3.1 Predefined static gestures recognized by the system during tests

Three sets of samples within the motion trajectories were used to form training and validation sets, split in the proportion of 66.7% to 33.3%, respectively. Two-fold cross-validation method was used for the first stage rough efficacy evaluation. For all persons, half of the sets containing the first motion trajectory samples formed a validating set and half of the sets consisting of the samples of the two remaining trajectories were used as a training set. Such a division provided a high discrimination between recognition efficacies for the classifiers and significantly decreased computational cost in comparison to the classical *n*-fold cross-validation or leave-one-out method.

In order to test the classifier efficacy and efficiency, an application in Java language based on the WEKA system classes was engineered. Parameters providing the highest performance for each classifier were found using a coarse grid-search method [11] with a step for each value equal to 2^k, $k \in [-m, n]$, m, n being integers chosen respectively for a particular parameter. The ranges for k were set basing on the review of the literature and empirical rough parameter adjustment. The step for k case equaled 1. Therefore, for the NNge classifier the number of folders for mutual information i, and the number of attempts for generalization g, took values from a range $[2^0, 2^4]$. The Naive Bayes classifier was examined for three conditions, i.e. using a kernel estimation for numeric attributes, using normal distribution of the numeric attributes and using supervised discretization to convert the numeric attributes to nominal ones. For J48 which was the WEKA implementation of the C4.5 classifier, the confidence factor C used for pruning took values from a range $[2^{-17}, 2^0]$ and the minimal number of instances per leaf m contained in range $[2^1, 2^6]$. For the Random Tree classifier, the number of randomly chosen attributes k_{rt} took values from a range $[2^0, 2^{10}] \cup \{\log_2 n_{rt} + 1\}$, where n_{rt} was the number of attributes, and the minimum total weight of the instances in leaf m was from a range $[2^{-17}, 2^0]$. For the Random Forest classifier, the number of trees took values from a range $[2^0, 2^{10}]$ and the number of attributes in random selection k_{rf} was from a range $[2^0, 2^5] \cup \{\log_2 n_{rf} + 1\}$, where n_{rf} was the number of attributes. For Multilayer Perceptron, learning rate value l and momentum m applied to weights during updating were selected from a range $[2^{-4}, 2^0]$ and the number of epochs to train through e contained within the range $[2^2, 2^5]$. For SVM, cost function C was selected from range $[2^0, 2^{14}]$ and *gamma* took values from range $[2^{-15}, 2^{-5}]$. These ranges were used both for linear and RBF kernels.

8.3.2 Results

The results of classifiers examination for the parameters providing the best performance are presented in Tables 8.3.1 and 8.3.2. The results have been ordered by performance (ascending).

Table 8.3.1 Results of classifiers examination for left hand

Classifier name	Average efficacy [%]	Average training time [ms]	Average validation time [ms]	Parameters
Random Tree	77.04	443	3	k = 2^6, m = 2^{-17}
C4.5 (J48)	77.73	1342	4	C = 2^{-7}, m = 2
Naive Bayes	79.49	303	73	supervised discretization
NNge	83.47	14234	8073	g = 2^2, i = 2^4
Random Forest	89.91	59644	722	i = 2^9, k = 2^4, unlimited depth

ANN (Multilayer Perceptron)	91.67	1458	187	l = 2^{-3}, m = 2^{-5},
				e = 2^3, one hidden layer, 4 nodes

SVM (LibSVM, linear kernel)	92.82	2599	1123	$\gamma = 2^{-15}$, C = 2^1
SVM (LibSVM, RBF kernel)	92.82	2508	1159	$\gamma = 2^{-11}$, C = 2^{11}

Table 8.3.2 Results of classifiers examination for right hand

Classifier name	Average efficacy [%]	Average training time [ms]	Average validation time [ms]	Parameters
NNge	75.42	8610	5409	g = 2^3, i = 2^2
C4.5 (J48)	76.44	1292	4	C = 2^{-6}, m = 2^3
Random Tree	77.92	993	4	k = 2^8, m = 2^{-17}
Naive Bayes	78.56	189	5231	using kernel estimation
Random Forest	84.44	15047	112	i = 2^7, k = 2^4, unlimited depth
SVM (LibSVM, linear kernel)	88.52	2811	1430	$\gamma = 2^{-8}$, C = 2^2
SVM (LibSVM, RBF kernel)	88.52	5166	2194	$\gamma = 2^{-6}$, C = 2^{12}

ANN (Multilayer Perceptron)	91.11	3854	181	l = 2^{-3}, m = 2^{-3},
				e = 2^5, one hidden layer, 4 nodes

The Multilayer Perceptron and SVM classifier have provided the highest recognition efficacy. Moreover, for different C and γ parameters, no matter whether a linear or RBF kernel was used, the efficacy was the same. This is consistent with findings in literature according to which for particular sets of C and γ parameters RBF kernel ensures at least such a performance as linear kernel [11].

The differences between the efficacies obtained for the left and right hand are relevant to the phenomenon already described in the literature (Kupryjanow *et al.*, 2010). This concerns right-handed persons, who were in the majority among the group tested. They obtained higher classification efficacy for the left hand drawing in the air, which is probably due to paying more attention to this activity, thus drawing more distinctive and easier distinguishable shapes for the classification system.

For the SVM and Multilayer Perceptron the leave-one-out cross-validation method was additionally used for testing. A validation set was a set of samples representing a particular motion trajectory of a particular person. Sets of samples within two other trajectories for other persons constituted a training set. Such a method of testing had an advantage over a typical cross-validation with random partitioning. It provided the possibility to examine the generalization ability of a classifier. Thus, it could be determined whether a predefined training set can be used in the system or an individual calibration aiming at collecting training data should be performed by the user. In tables 8.3.3 – 8.3.5 the results of leave-one-out validation for SVM and ANN classifiers are presented.

Table 8.3.3 Results of leave-one-out cross-validation for LibSVM with linear kernel

	Left Hand	Right hand
Min. efficacy [%]	66.67	65.83
Max. efficacy [%]	100.00	100.00
Average efficacy [%]	95.68	94.65
Median [%]	98.33	97.50
Unbiased standard deviation	6.43	7.82
Unbiased variance	40.15	59.34
Low limit of 95% confidence interval	93.85	92.42
High limit of 95% confidence interval	97.51	96.87
Skewness	-2.30	-2.07
Kurtosis	6.14	4.12
Average training time [ms]	6435	6598
Average validation time [ms]	197	203

High median, high variance and negative skewness suggest that there was a person or a few persons who performed gestures significantly differently than the majority. Therefore, the assumption for the final system has been made that individual calibration phase for a particular user should be performed instead of using a predefined training set.

Table 8.3.4 Results of leave-one-out cross-validation for LibSVM with RBF kernel

	Left Hand	Right hand
Min. efficacy [%]	65.83	65.83
Max. efficacy [%]	100.00	100.00
Average efficacy [%]	95.56	94.09
Median [%]	98.33	98.33
Unbiased standard deviation	6.59	7.97
Unbiased variance	42.16	61.75
Low limit of 95% confidence interval	93.68	91.82
High limit of 95% confidence interval	97.43	96.36
Skewness	-2.28	-1.47
Kurtosis	6.03	1.50
Average training time [ms]	6575	10478
Average validation time [ms]	204	335

Table 8.3.5 Results of leave-one-out cross-validation for ANN

	Left Hand	Right hand
Min. efficacy [%]	60.83	69.17
Max. efficacy [%]	100.00	100.00
Average efficacy [%]	92.53	93.24
Median [%]	96.25	95.83
Unbiased standard deviation	7.99	7.36
Unbiased variance	61.79	52.64
Low limit of 95% confidence interval	90.26	91.15
High limit of 95% confidence interval	94.80	95.33
Skewness	-1.56	-1.16
Kurtosis	2.84	0.68
Average training time [ms]	2386	2379
Average validation time [ms]	21	21

For the LibSVM with a linear kernel we additionally performed a fine grid-search to adjust some selected parameters. According to known research (Hsu *et al.*, 2010), performing fine grid-search allows to find parameters for which the overall performance of the classifier can be increased by 0.1%. It was found that the best parameters are: $C = 2^{1.25}$, $\gamma = 2^{-15}$ for the left hand, and $C = 2^{2}$, $\gamma = 2^{-8.75}$ for the right hand, providing the increase of the average efficacy by 0.1% in the first case and no increase in the second case.

8.4 Conclusions

The highest efficacy of static gestures recognition was obtained for SVM and ANN. The next best performing classifier was Random Forest. However, the last mentioned classifier was rejected due to unacceptably long training time. Despite similar performance of SVM and ANN, and shorter validation time of the latter, we decided to choose the first one for the engineered system. The ANN classifier was rejected due to the risk of performance decrease being the effect of overtraining in case of changing the size of training set, at the stage of the software maintenance. Moreover, the minimal efficacy for ANN was lower than for the SVM and the ANN did not guarantee repeatability of classification results for identical shapes.

Considering research carried out on similar problems, reported in literature [13, 19], the high performance of SVM applied to shape recognition is not surprising. Due to its performance and predictability this particular classifier is often the first choice for solving vast array of problems. However, recently the Optimum-Path Forest classifier, which in many cases outperforms SVM, has been gaining popularity [1, 17, 18]. Using the OPF classifier in the engineered system instead of SVM could further improve gestures recognition and significantly decrease training time.

Acknowledgments. Research funded within the project No. POIG.01.03.01-22-017/08, entitled "Elaboration of a series of multimodal interfaces and their implementation to educational, medical, security and industrial applications". The project is subsidized by the European regional development fund and by the Polish State budget.

References

1. Albuquerque, V.H.C., et al.: Application of Optimum-Path Forest Classifier for Synthetic Material Porosity Segmentation. In: 17th International Conference on Systems, Signals and Image Processing (IWSSIP 2010), pp. 1–4 (2010)
2. Athitsos, V., Sclaroff, S.: Estimating 3D Hand Pose from a Cluttered Image. In: IEEE Conference on Computer Vision and Pattern Recognition, Wisconsin, pp. 432–439 (2003)
3. Bajaj, C.L., Pascucci, V., Schikore, D.R.: The contour spectrum. IEEE Visualization, 167–173 (1997)
4. Beck, J.R., Garcia, M., Zhong, M., Georgiopoulos, M.: A Backward Adjusting Strategy and Optimization of the C4.5 Parameters to Improve C4.5's Performance. In: XXI Artificial Intelligence Research Symposium FLAIRS, Coconut Grove, FL, pp. 35–40 (2008)
5. Bosch, A., Zisserman, A., Munoz, X.: Image Classification using Random Forests and Ferns. In: 11th International Conf. on Computer Vision ICCV, pp. 1–8. IEEE, Rio de Janeiro (2007)
6. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly, Sebastopol (2008)
7. Brent, M.: Instance-based Learning: Nearest Neighbour with Generalisation. Hamilton, New Zealand, M.Sc. thesis, University of Waikato (1995)
8. Carr, H., Snoeyink, J., van de Panne, M.: Progressive topological simplification using contour trees and local spatial measures. In: 15th Western Computer Graphics Symposium, British Columbia (2004)

9. Chang, C.C., Lin, C.J.: LIBSVM: a Library for Support Vector Machines. Science, 1–39 (2011), http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed October 27, 2011)
10. Holmes, G., Donin, A., Witten, I.: WEKA: A Machine Learning Workbench. In: 2nd Australian and New Zealand Conf. on Intelligent Inform. Systems, pp. 357–361. IEEE, Brisbane (1994)
11. Hsu, C.W., Chang, C.C., Lin, C.J.: A Practical Guide to Support Vector Classification. Bioinformatics 1(1), 1–16 (2010), http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed October 27, 2011)
12. Krasser, S., et al.: Identifying Image Spam based on Header and File Properties using C4.5 Decision. In: Proceedings of the 2007 IEEE Workshop on Information Assurance, pp. 255–261. IEEE, New York (2007)
13. Kupryjanow, A., Kaszuba, K., Czyzewski, A.: Influence of accelerometer signal preprocessing and classification method on human activity recognition. Elektronika 51(3), 18–23 (2010)
14. Lech, M., Kostek, B.: Fuzzy Rule-Based Dynamic Gesture Recognition Employing Camera and Multimedia Projector. In: Nguyen, N.T., Zgrzywa, A., Czyżewski, A. (eds.) Advances in Multimedia and Network Information System Technologies. AISC, vol. 80, pp. 69–78. Springer, Heidelberg (2010)
15. Lech, M., Kostek, B.: Hand Gesture Recognition Supported by Fuzzy Rules and Kalman Filters. International Journal of Intelligent Information and Database Systems 6(5) (2012) (forthcoming paper)
16. Mitra, S., Acharya, T.: Gesture Recognition: A Survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 311–324 (2007)
17. Papa, J.P., et al.: Design of robust pattern classifiers based on optimum-path forests. In: 8th International Symposium on Mathematical Morphology, MCT/INPE, Rio de Janeiro, pp. 337–348 (2007)
18. Papa, J.P., Falcao, A.X.: Supervised Pattern Classification based on Optimum-Path Forest. Int. Journal of Imaging Systems and Technology 19(2), 120–131 (2009)
19. Pumpuang, P., Srivihok, A., Praneetpolgrang, P.: Comparisons of Classifier Algorithms: Bayesian Network, C4.5, Decision Forest and NBTree for Course Registration Planning Model of Undergraduate Students. In: IEEE International Conference on Systems, Man and Cybernetics (SMC 2008), pp. 3647–3651. IEEE, Singapore (2008)
20. Reeb, G.: Sur les points singuliers d'une forme de Pfaff completement integrable ou d'une fonction numerique. Comptes Rendus de l'Academie des Sciences 222, 847–849 (1946)
21. Sang-Bum, K., Kyoung-Soo, H., Hae-Chang, R., Sung Hyon, M.: Some Effective Techniques for Naive Bayes. IEEE Transactions on Knowledge and Data Engineering 18(11), 1457–1466 (2006)
22. Vlaardingen, M.: Hand Models and Systems for Hand Detection, Shape Recognition and Pose Estimation in Video. In: Pattern Recognition & Bioinformatics Group. Research Assignment, Delft Univ. of Technology, pp. 1–41 (2006)
23. Wang, W., Pollak, I., Bouman, C.A., Harper, M.P.: Classification of Images Using Spatial Random Trees. In: IEEE/SP 13th Workshop on Statistical Signal Processing, pp. 449–452. IEEE, Novosibirsk (2005)
24. Xiuxin, Y., Anh, D., Li, C.: A wearable real-time fall detector based on Naive Bayes classifier. In: 23rd Canadian Conf. on Electrical and Computer Engineering (CCECE), pp. 1–4. IEEE (2010)

Part II
Information Systems Specification

Chapter 9
Multiple Data Tables Processing via One-Sided Concept Lattices

Peter Butka, Jozef Pócs, Jana Pócsová, and Martin Sarnovský

Abstract. One-sided concept lattices introduce data mining method from the area of Formal Concept Analysis (FCA) for analysis of objects clusters according to the set of fuzzy attributes. Currently, most of the methods for creation of one-sided concept lattices process only data tables with one type of truth value structure. In this chapter we describe closure operator, which corresponds to the intersection of particular closure systems obtained from various object-attribute models with different types of attributes. Each particular closure system is defined via one-sided concept lattices approaches applicated for particular data tables.

9.1 Introduction

Formal Concept Analysis (FCA, [5]) is a theory of data analysis for identification of conceptual structures among data sets. FCA has been found useful in concept data analysis, knowledge discovery, text mining, information retrieval, as well as in other areas from machine learning and artificial intelligence. Classic approach to FCA provides crisp case, where object-attribute model is based on the binary relation (object has/has-not the attribute). However, in practice there are natural examples of object-attribute models for which relationship between objects and attributes are

Peter Butka · Martin Sarnovský
Technical University of Košice, Faculty of Electrical Engineering and Informatics,
Department of Cybernetics and Artificial Intelligence, Letná 9, 04200 Košice, Slovakia
e-mail: peter.butka@tuke.sk, martin.sarnovsky@tuke.sk

Jozef Pócs
Mathematical Institute, Slovak Academy of Sciences, Grešákova 6, 040 01 Košice, Slovakia
e-mail: pocs@saske.sk

Jana Pócsová
Technical University of Košice, BERG Faculty, Institute of Control and Informatization
of Production Processes, Boženy Němcovej 3, 043 84 Košice, Slovakia
e-mail: jana.pocsova@tuke.sk

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 89–98.
springerlink.com

represented by fuzzy relations. Therefore, several attempts to fuzzify FCA have been proposed. We mention an approach of Bělohlávek [3, 4], an approach of Krajči [10], Popescu [16], other approaches [13, 15, 17], and also work on fuzzy FCA in categorical settings [11, 12].

Important role in fuzzy FCA play one-sided concept lattices, where usually objects are considered as a crisp subsets and attributes obtain fuzzy values. In this case on the side of objects are crisp sets and on the side of attributes there are fuzzy sets. From existing one-sided approaches we mention papers of Krajči [9], Yahia and Jaoua [2], work of Jaoua and Elloumi on Galois lattices of real relations [8]. In case of one-sided concept lattices there is strong connection with clustering (cf. [7]). As it is known, clustering methods produce subsets of a given set of objects, which are closed under intersection, i.e., closure system on the set of objects. Since one-sided concept lattice approach produces also closure system on the set of objects, one can see one-sided concept lattice approaches as a special case of hierarchical clustering.

The aim of this chapter is to introduce the method for processing of multiple data tables in case, where each data table describe the fuzzy relation between the set of objects and different set of attributes. The entries of each data tables are values from different truth value structure corresponding to given set of attributes. It is logical for practice that various types of attributes are represented by different values, e.g., binary attributes with possible values $0, 1$, or real-valued attributes with values from real interval. Above mentioned methods for creation of one-sided concept lattices are not able to process data table with different types of attributes. One of the possible ways is to analyze input data table as several data tables, where each table contains only attributes of same type. Hence, it is possible to create closure system for every data table and consequently produce the intersection of them. The resulting closure system is composed from the clusters which are contained in all particular closure systems, therefore these clusters are significant for each object-attribute model. Hence, the main goal is to mathematically describe resulting closure operator, corresponding to closure system, defined as intersection of closure systems given by particular concept lattices.

In the following section we provide necessary algebraic preliminaries for description of one-sided concept lattices, i.e., Galois connections, closure operators, closure systems, etc. Section 9.3 is devoted to closure operator defined by multiple data tables. We provide and prove our basic results, i.e., description of closure operator corresponding to the intersection of closure systems defined for multiple data tables.

9.2 Preliminaries on One-Sided Concept Lattices

In this section we describe fuzzy generalization of classical concept lattices, so called one-sided concept lattices.

The main idea of fuzzifications of classical FCA is the usage of graded truth. In classical logic, each proposition is either true or false, hence classical logic is bivalent. In fuzzy logic, to each proposition there is assigned a truth degree from some scale. The underlying structure of truth degrees is partially ordered and contains the smallest and the greatest element. If to the propositions ϕ and ψ are assigned truth degrees $\parallel \phi \parallel = a$ and $\parallel \psi \parallel = b$, then $a \leq b$ means that ϕ is considered less true than ψ. In object-attribute models the typical propositions are of the form "object has attribute in degree a". The structures of truth degrees commonly used in various modifications of fuzzy logic are real unit interval $[0,1]$, Boolean algebras, MV-algebras or more generally residuated lattices. All this structures are equipped with binary operations simulating implication and the logical connective and, but the important fact is that they form a complete lattice according to the partial order defined on them. In order to introduce the notion of one-sided concept lattices as a generalization of FCA we will assume only one minimal condition, i.e., the structures of truth degree forms complete lattice. In what follows we will assume that the reader is familiar with the basic notions of lattice theory (see [6]).

Crucial role in the mathematical theory of fuzzy concept lattices play special pairs of mappings between complete lattices, commonly known as Galois connections. Hence, we provide necessary details regarding Galois connections and related topics.

Let $(P,\leq)$ and $(Q,\leq)$ be complete lattices and let $\varphi\colon P \to Q$ and $\psi\colon Q \to P$ be maps between these lattices. Such a pair (φ,ψ) of mappings is called a *Galois connection* if the following condition is fulfilled:

$$p \leq \psi(q) \quad \text{if and only if} \quad \varphi(p) \geq q.$$

Galois connections between complete lattices are closely related to the notion of closure operator and closure system. Let L be a complete lattice. By a *closure operator* in L we understand a mapping $c\colon L \to L$ satisfying:

(a) $x \leq c(x)$ for all $x \in L$,
(b) $c(x_1) \leq c(x_2)$ for $x_1 \leq x_2$,
(c) $c(c(x)) = c(x)$ for all $x \in L$ (i.e., c is idempotent).

Subset X of the complete lattice L is called *closure system* in L if X is closed under arbitrary meets. We note, that this condition guarantees that $(X,\leq)$ is a complete lattice, in which the infima are the same as in L, but the suprema in X may not coincide with those from L. For a closure operator c in L, the set $\mathsf{FP}(c)$ of all fixed points of c (i.e., $\mathsf{FP}(c) = \{x \in L : c(x) = x\}$) is the closure system in L. Conversely, for closure system X in L, mapping $\mathsf{C}_X : L \to L$ defined by $\mathsf{C}_X(x) = \bigwedge\{u \in X : x \leq u\}$ is the closure operator in L. Moreover these correspondences are inverses of each other, i.e., $\mathsf{FP}(\mathsf{C}_X) = X$ for each closure system X in L and $\mathsf{C}_{\mathsf{FP}(c)} = c$ for each closure operator c in L.

First, we describe Galois connections between power sets, which are the cornerstones of the classical FCA.

Let (B,A,I) be a formal context, i.e., $B,A \neq \emptyset$ and $I \subseteq B \times A$ be a binary relation between B and A. There is defined a pair of mappings $^{\uparrow} : 2^B \to 2^A$ and $^{\downarrow} : 2^A \to 2^B$ as follows:

$$X^{\uparrow} = \{y \in A : (x,y) \in I \text{ for all } x \in X\},$$

$$Y^{\downarrow} = \{x \in B : (x,y) \in I \text{ for all } y \in Y\}.$$

As it can be easily shown, this pair of mappings forms a Galois connection. Based on this mappings there is defined a complete lattice called *concept lattice*, c.f. [5] for detail mathematical description.

One-sided concept lattices are defined via Galois connection between 2^B (set of all subsets of a given set of objects) and the powers of arbitrary complete lattices L^A, which represent L- fuzzy subsets. In what follows, we describe a method of constructing one-side concept lattices from the given formal context.

Formally, let $B \neq \emptyset$ be the set of objects, $A \neq \emptyset$ be the set of attributes and $R : B \times A \to L$ be L-fuzzy relation. A triple $\big(B,A,R\big)$ is said to be L *one-sided formal context*. The value $R(b,a)$ represents a degree from the structure of truth values L in which the element $b \in B$ has the attribute a.

Now we provide a definition of a pair of mappings used for construction of one-sided concept lattices, see [9].

Let $\big(B,A,R\big)$ be one-sided formal context. Then there is defined a pair of mapping ${}^{\triangle} : 2^B \to L^A$ and ${}^{\triangledown} : L^A \to 2^B$ as follows:

$$X^{\triangle}(a) = \bigwedge_{b \in X} R(b,a),$$

$$g^{\triangledown} = \{b \in B : \text{ for each } a \in A,\ g(a) \leq R(b,a)\}.$$

The pair $({}^{\triangle}, {}^{\triangledown})$ forms a Galois connection between 2^B and L^A. The composition of mappings ${}^{\triangle}$ and ${}^{\triangledown}$ forms closure operator in 2^B and similarly the composition of ${}^{\triangledown}$ and ${}^{\triangle}$ forms closure operator in L^A. Hence, subsets of the form $X^{\triangle\triangledown}$ for any $X \subseteq B$ are closed subsets with respect to the closure operator defined above. As it is known fact, the closed subsets of any closure operator forms a complete lattice, with respect to the inherited partial order from underlying complete lattice structure (in this case 2^B). This fact stands behind the formal definition and characterization of concept lattices.

For a given formal context $\big(B,A,R\big)$ the symbol $\mathscr{C}\big(B,A,R\big)$ will denote the set of all pairs (X,g), where $X \subseteq B$, $g \in L^A$, satisfying

$$X^{\triangle} = g \quad \text{and} \quad X = g^{\triangledown}.$$

In this case, the set X is usually referred as *extent* and g as *intent* of the concept (X,g). Further we define partial order on $\mathscr{C}\big(B,A,\mathsf{L},R\big)$ as follows:

$$(X_1,g_1) \leq (X_2,g_2) \text{ iff } X_1 \subseteq X_2 \text{ iff } g_1 \geq g_2.$$

Proposition 1. *Let $\big(B,A,R\big)$ be one-sided formal context. The set $\mathscr{C}\big(B,A,R\big)$ with the partial order defined above forms a complete lattice, where*

$$\bigwedge_{i \in I}(X_i,g_i) = \Big(\bigcap_{i \in I} X_i, \big(\bigvee_{i \in I} g_i\big)^{\triangledown\triangle}\Big) = \Big(\big(\bigvee_{i \in I} g_i\big)^{\triangledown}, \big(\bigcap_{i \in I} X_i\big)^{\triangle}\Big)$$

and

$$\bigvee_{i\in I}(X_i,g_i)=\Big(\big(\bigcup_{i\in I}X_i\big)^{\triangle\nabla},\bigwedge_{i\in I}g_i\Big)=\Big(\big(\bigwedge_{i\in I}g_i\big)^{\nabla},\big(\bigcup_{i\in I}X_i\big)^{\triangle}\Big)$$

for each family $(X_i,g_i)_{i\in I}$ *of elements from* $\mathscr{C}\big(B,A,R\big)$.

The lattice $\mathscr{C}\big(B,A,R\big)$ is called *one-sided concept lattices*.

9.3 Closure Operator Defined by Multiple Data Tables

Although the theory of one-sided concept lattices described in [9] does not allow to deal with different truth value structures, it is reasonable to consider similar properties (in our case closure system) on the side of objects defined using various data tables, even if this closure system is not directly represented as a part of one-sided concept lattices. One of the possibility is to define corresponding closure system as the intersection of all closure systems obtained by Galois connections from the given formal contexts. In this case, the resulting closure system will consist from subset of objects which are precisely contained in all particular concept lattices. The main aim of this section is to introduce resulting closure system and to describe the corresponding closure operator.

Let $B\neq\emptyset$ be a set of objects and $\{A_i\}_{i=1}^{k}$ be a system of attributes, where each attribute a of the set A_i is represented by truth value structure L_i. Further, we will assume that the relationships between the set of objects B and particular attributes A_i are described by L_i-fuzzy relation $R_i\colon B\times A_i\to L_i$. Galois connections corresponding to the generalized one-sided formal contexts (B,A_i,R_i) for $i=1,\ldots,k$, will be denoted by $({}^{\triangle_i},{}^{\nabla_i})$. Further we denote by $\mathsf{FP}(i)$ the set of all fixed points of the closure operator induced by the composition of mappings ${}^{\triangle_i}$ and ${}^{\nabla_i}$.

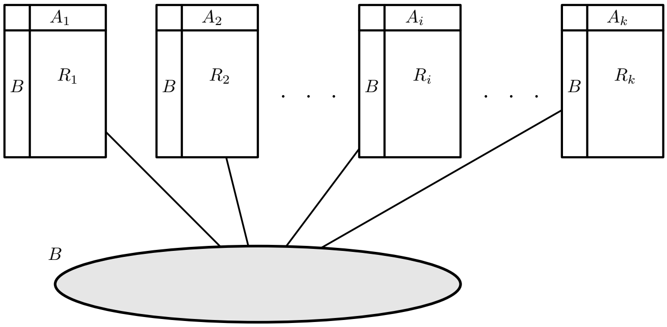


Fig. 9.1 Object-attribute models connected with the same set of objects

We define a closure system $\triangle \subseteq 2^B$ as the intersection of closure systems $\mathsf{FP}(i)$, i.e.,

$$\triangle = \bigcap_{i=1}^{k} \mathsf{FP}(i).$$

As one can easily see, such defined system of subsets forms a closure system, which is the greatest closure system contained in each $\mathsf{FP}(i)$ for $i = 1,\dots,k$.

Using Galois connections $({}^{\triangle_i}, {}^{\nabla_i})$ we define an operator ${}^{\diamondsuit}$ on the set B of all objects, i.e., ${}^{\diamondsuit}\colon 2^B \to 2^B$ with the property $\triangle = \mathsf{FP}({}^{\diamondsuit})$. First, for any $X \subseteq B$ we define by induction a sequence of subsets $\{Y_j\}_{j=0}^{\infty}$ as follows:

$$Y_0 = X, \quad Y_{j+1} = \bigcup_{i=1}^{k} Y_j{}^{\triangle_i \nabla_i}, \quad \text{for all } j \geq 0.$$

Lemma 1. *Let $B \neq \emptyset$ be a finite set of objects. Then there exists an index $j \in \mathbb{N}_0$ such that $Y_j = Y_l$ for any $l \geq j$.*

Proof. Let us suppose that B contains n elements. First, observe that $Y_j \subseteq Y_{j+1}$ for any $j \in \mathbb{N}_0$. This follows from the fact that $Y_j \subseteq Y_j{}^{\triangle_i \nabla_i}$ for all $i = 1,\dots,k$, hence applying the union operation one obtain $Y_j \subseteq \bigcup_{i=1}^{k} Y_j{}^{\triangle_i \nabla_i} = Y_{j+1}$. Consequently, there is a nondecreasing sequence

$$|Y_0| \leq |Y_1| \leq |Y_2| \leq \dots \leq \left|Y_j\right| \leq \dots \leq n$$

of nonnegative numbers, which is bounded above by cardinality of the set B. Since B is finite, there exists an index j such that $\left|Y_j\right| = \left|Y_{j+1}\right|$ what implies $Y_j = Y_{j+1}$. Obviously, from the definition of the sequence $\{Y_j\}_{j=0}^{\infty}$ we obtain $Y_j = Y_l$ for all $l \geq j$. □

Let us note that the similar assertion can be also obtained for infinite object sets using Tarski's fixed point theorem. Using the result of Lemma 1 we can correctly define the operator ${}^{\diamondsuit}$ as follows:

$$X^{\diamondsuit} = Y_j, \text{ for smallest } j \text{ with } Y_j = Y_{j+1}.$$

Theorem 1. *The operator ${}^{\diamondsuit}\colon 2^B \to 2^B$ forms a closure operator on the set of objects B, moreover $\triangle = \mathsf{FP}({}^{\diamondsuit})$.*

Proof. We show that operator ${}^{\diamondsuit}$ satisfies the properties of closure operator.

(a) $X \subseteq X^{\diamondsuit}$ for all $X \subseteq B$. This follows from the fact that $X = Y_0 \subseteq Y_j$ for all $j \geq 0$. Thus $X \subseteq X^{\diamondsuit} = Y_j$ for minimal j with $Y_j = Y_{j+1}$.

(b) $X_1{}^{\diamondsuit} \subseteq X_2{}^{\diamondsuit}$ for $X_1 \subseteq X_2$. Let $\{Y_j^1\}_{j=0}^{\infty}$, $\{Y_j^2\}_{j=0}^{\infty}$ be sequences corresponding to subsets X_1 and X_2 respectively. Using mathematical induction we prove $Y_j^1 \subseteq Y_j^2$ for all $j \geq 0$. For $j = 0$ this is obvious, since $Y_0^1 = X_1 \subseteq X_2 = Y_0^2$. Assume that $Y_j^1 \subseteq Y_j^2$. Then $Y_j^{1\,\triangle_i \nabla_i} \subseteq Y_j^{2\,\triangle_i \nabla_i}$ for all $i = 1,\dots,k$, which implies

$$Y_{j+1}^{1}=\bigcup_{i=1}^{k}Y_{j}^{1\,\triangle_i\nabla_i}\subseteq\bigcup_{i=1}^{k}Y_{j}^{2\,\triangle_i\nabla_i}=Y_{j+1}^{2}.$$

Further, suppose that ${X_1}^{\diamondsuit}=Y_r^1$ for some $r\geq 0$ and ${X_2}^{\diamondsuit}=Y_s^2$ for some $s\geq 0$. According to Lemma 1 and above inequality $Y_j^1\subseteq Y_j^2$ we obtain

$${X_1}^{\diamondsuit}=Y_{\max\{r,s\}}^{1}\subseteq Y_{\max\{r,s\}}^{2}={X_2}^{\diamondsuit}.$$

(c) $X^{\diamondsuit\diamondsuit}=X^{\diamondsuit}$ for all $X\subseteq B$. Let $X^{\diamondsuit}=Y_j$ for some $j\geq 0$. Then $X^{\diamondsuit\diamondsuit}={Y_j}^{\diamondsuit}$ and since $Y_{j+1}=Y_j$ we obtain ${Y_j}^{\diamondsuit}=Y_j$.

Hence, the operator $^{\diamondsuit}$ satisfies the properties of closure operators.

Next, we prove that $\triangle=\mathsf{FP}(^{\diamondsuit})$. First, we show that $\mathsf{FP}(^{\diamondsuit})\subseteq\mathsf{FP}(i)$ for all $i=1,\ldots,k$. Let $X\subseteq B$ be any fixed point of the operator $^{\diamondsuit}$, i.e., $X\in\mathsf{FP}(^{\diamondsuit})$. Then, due to the definition of the operator $^{\diamondsuit}$ and the fact that the least index j satisfying $Y_j=Y_{j+1}$ is equal to zero, we obtain

$$X=X^{\diamondsuit}=Y_0=Y_1=\bigcup_{i=1}^{k}{Y_0}^{\triangle_i\nabla_i}=\bigcup_{i=1}^{k}X^{\triangle_i\nabla_i}.$$

Since $X\subseteq X^{\triangle_i\nabla_i}$ for any $i=1,\ldots,k$, equality $X=\bigcup_{i=1}^{k}X^{\triangle_i\nabla_i}$ implies $X=X^{\triangle_i\nabla_i}$ for each $i=1,\ldots,k$. Hence, we have proved that any $X\in\mathsf{FP}(^{\diamondsuit})$ is also in the closure system $\mathsf{FP}(i)$, which implies $\mathsf{FP}(^{\diamondsuit})\subseteq\mathsf{FP}(i)$ for each $i=1,\ldots,k$ and we have $\mathsf{FP}(^{\diamondsuit})\subseteq\bigcap_{i=1}^{k}\mathsf{FP}(i)=\triangle$.

Conversely, we show $\triangle\subseteq\mathsf{FP}(^{\diamondsuit})$. Suppose that $X\in\triangle$, i.e., $X=X^{\triangle_i\nabla_i}$ for all $i=1,\ldots,k$. From this assumption we obtain

$$Y_1=\bigcup_{i=1}^{k}{Y_0}^{\triangle_i\nabla_i}=\bigcup_{i=1}^{k}X^{\triangle_i\nabla_i}=\bigcup_{i=1}^{k}X=X=Y_0,$$

thus the last index that stabilizes the sequence is equal zero and consequently from the definition of the operator $^{\diamondsuit}$ we obtain $X=X^{\diamondsuit}$. This equality shows that X is also fixed point of the operator $^{\diamondsuit}$, hence $\triangle=\mathsf{FP}(^{\diamondsuit})$, which completes the proof. □

Using closure system $\triangle$ we can formally define complete lattice $\mathscr{D}(\triangle)$ consisting of pairs $(X,\mathscr{M})$ where $X\subseteq B$ and $\mathscr{M}$ is k-tuple $(g_1,g_2,\ldots,g_k)$ such that each i-th component is L-fuzzy subset of A_i, i.e., $g_i\in L^{A_i}$, moreover $X^{\triangle_i}=g_i$ and $X={g_i}^{\nabla_i}$ for all $i=1,\ldots,k$. Partial order is defined by subset relation on power set 2^B, i.e.,

$$(X_1,\mathscr{M}_1)\leq(X_2,\mathscr{M}_2)\ \text{ iff }\ X_1\subseteq X_2. \tag{9.1}$$

According to the fact that $\triangle$ forms closure system and Theorem 1 we are able to provide the following characterization.

Theorem 2. *The set $\mathscr{D}(\triangle)$ with the partial order defined by (9.1) forms a complete lattice, with*

$$\bigwedge_{i\in I}(X_i,\mathscr{M}_i) = \Big(\bigcap_{i\in I} X_i, \underline{\mathscr{M}}\Big), \textit{ where } \underline{\mathscr{M}} = \Big((\bigcap_{i\in I} X_i)^{\triangle_1},\dots,(\bigcap_{i\in I} X_i)^{\triangle_k}\Big),$$

$$\bigvee_{i\in I}(X_i,\mathscr{M}_i) = \Big((\bigcup_{i\in I} X_i)^{\diamondsuit},\overline{\mathscr{M}}\Big), \textit{ where } \overline{\mathscr{M}} = \Big((\bigcup_{i\in I} X_i)^{\diamondsuit\triangle_1},\dots,(\bigcup_{i\in I} X_i)^{\diamondsuit\triangle_k}\Big)$$

for each family $\big(X_i,\mathscr{M}_i\big)_{i\in I}$ *of elements from* $\mathscr{D}(\triangle)$.

The proof of this theorem is similar with the proof for one-sided concept lattices (see [9]).

Next, we provide an example of the presented approach of generating closure system $\triangle$ from given one-sided contexts. We will consider five element set of objects $B = \{b_1,b_2,b_3,b_4,b_5\}$ and three formal contexts, where $A_1 = \{a_1,a_2,a_3,a_4\}$ and $L_1 = \mathbf{4}$ represents four element ordinal scale with values $0 \le 1 \le 2 \le 3$, $A_2 = \{a_5,a_6,a_7,a_8,a_9\}$ and $L_2 = \mathbf{2}$ is two element chain, $A_3 = \{a_{10},a_{12},a_{12}\}$ and $L_3 = [0,1]$ is real unit interval. The L_i-fuzzy relations R_i for $i = 1,2,3$ are depicted in the following tables.

Table 9.1 Example of three data tables corresponding to relations R_1,R_2,R_3

R_1	a_1	a_2	a_3	a_4
b_1	2	0	1	3
b_2	3	1	0	2
b_3	1	2	0	1
b_4	3	0	1	0
b_5	0	2	1	1

R_2	a_5	a_6	a_7	a_8	a_9
b_1	0	1	1	1	1
b_2	1	0	0	1	0
b_3	1	1	1	0	0
b_4	0	1	0	1	1
b_5	1	0	1	1	1

R_3	a_{10}	a_{11}	a_{12}
b_1	0.4	0.6	0.3
b_2	0.3	0.2	0.1
b_3	0.1	0.7	0.5
b_4	0.2	0.4	0.6
b_5	0.5	0.2	0.7

Let us remark, that in this case R_2 represents ordinary binary relation and $R_2(b,a) = 1$ if and only if object b has attribute a.

Figure 9.2 shows the particular concept lattices corresponding to the relations R_1,R_2,R_3. Intersection of their corresponding closure systems leads to the concept lattice provided on Figure 9.3. As one can see, intersection significantly reduces the number of concepts (clusters) in the resulted model only to those, which are included in all particular concept lattices.

This approach can be used for various cases. For example, we can imagine the information retrieval system based on the intersection of object-attribute models defined for query with different input attributes. Also it is possible to apply it to any data mining problem related to analysis of clusters of objects and their relations, e.g., meteorological data or analysis of logs in e-learning domain, cf. [1, 14].

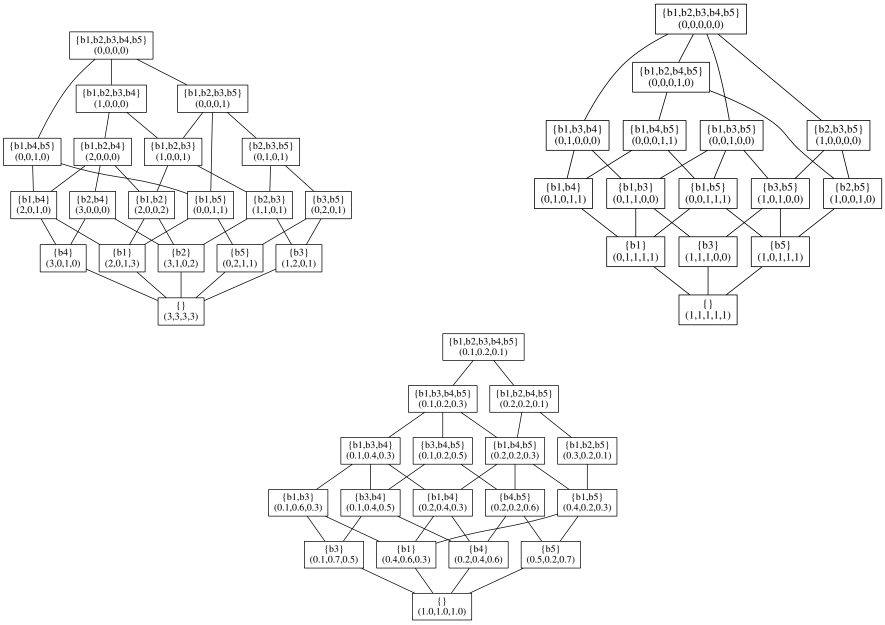


Fig. 9.2 Concept lattices corresponding to relations R_1 (left), R_2 (middle), R_3 (right)

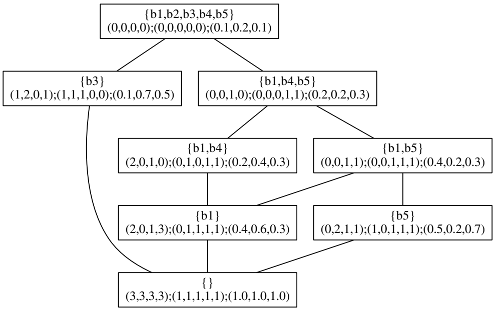

{b1,b2,b3,b4,b5} (0,0,0,0);(0,0,0,0,0);(0.1,0.2,0.1)
{b3} (1,2,0,1);(1,1,1,0,0);(0.1,0.7,0.5)
{b1,b4,b5} (0,0,1,0);(0,0,0,1,1);(0.2,0.2,0.3)
{b1,b4} (2,0,1,0);(0,1,0,1,1);(0.2,0.4,0.3)
{b1,b5} (0,0,1,1);(0,0,1,1,1);(0.4,0.2,0.3)
{b1} (2,0,1,3);(0,1,1,1,1);(0.4,0.6,0.3)
{b5} (0,2,1,1);(1,0,1,1,1);(0.5,0.2,0.7)
{} (3,3,3,3);(1,1,1,1,1);(1.0,1.0,1.0)

Fig. 9.3 Result structure $\mathscr{D}(\triangle)$ from the relations R_1, R_2, R_3

9.4 Conclusions

In this chapter we have presented the method for processing of multiple data tables with various truth value structures via different one-sided concept lattices approaches. If original data table contains several types of attributes, it is possible to create suitable one-sided concept lattice as an intersection of closure systems (which correspond to particular concept lattices) for subtables generated from the original one for every type of attribute. Therefore, most interesting subsets of objects (clusters) are those, which are contained in all particular lattices. Moreover, our approach leads to resulted concept lattice with significantly reduced size, what supports better and more precise visualization of the information contained in such types of conceptual models.

Acknowledgements. This work was supported by the Slovak Research and Development Agency under contracts APVV-0208-10 (40%) and APVV-0035-10 (10%), by the Development of the Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030)(30%), and also by the Slovak VEGA Grant No. 1/1147/12 (20%).

References

1. Babič, F., Bednár, P., Albert, F., Paralič, J., Bartók, J., Hluchý, L.: Meteorological Phenomena Forecast Using Data Mining Prediction Methods. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 458–467. Springer, Heidelberg (2011)
2. Ben Yahia, S., Jaoua, A.: Discovering knowledge from fuzzy concept lattice. In: Kandel, A., Last, M., Bunke, H. (eds.) Data Mining and Computational Intelligence, pp. 167–190. Physica-Verlag (2001)
3. Bělohlávek, R.: Lattices generated by binary fuzzy relations. Tatra Mt. Math. Publ. 16, 11–19 (1999)
4. Bělohlávek, R.: Lattices of Fixed Points of Fuzzy Galois Connections. Math. Log. Quart. 47(1), 111–116 (2001)
5. Ganter, B., Wille, R.: Formal concept analysis: Mathematical foundations. Springer, Berlin (1999)
6. Grätzer, G.: Lattice Theory: Foundation. Springer, Basel (2011)
7. Janowitz, M.F.: Ordinal and relational clustering. World Scientific Publishing Company, Hackensack (2010)
8. Jaoua, A., Elloumi, S.: Galois connection, formal concepts and Galois lattice in real relations: application in a real classifier. J. Syst. Software 60, 149–163 (2002)
9. Krajči, S.: Cluster based efficient generation of fuzzy concepts. Neural Netw. World 13(5), 521–530 (2003)
10. Krajči, S.: A generalized concept lattice. Logic Journal of IGPL 13(5), 543–550 (2005)
11. Krídlo, O., Ojeda-Aciego, M.: On L-fuzzy Chu correspondences. Int. J. Comput. Math. 88(9), 1808–1818 (2011)
12. Krídlo, O., Krajči, S., Ojeda-Aciego, M.: The Category of L-Chu Correspondences and the Structure of L-Bonds. Fund. Inform. 115(4), 297–325 (2012)
13. Medina, J., Ojeda-Aciego, M., Ruiz-Calviño, J.: Formal concept analysis via multi-adjoint concept lattices. Fuzzy Set. Syst. 160, 130–144 (2009)
14. Paralič, J., Richter, C., Babič, F., Wagner, J., Raček, M.: Mirroring of Knowledge Practices based on User-defined Patterns. J. Univers. Comput. Sci. 17(10), 1474–1491 (2011)
15. Pócs, J.: Note on generating fuzzy concept lattices via Galois connections. Inform. Sci. 185(1), 128–136 (2012)
16. Popescu, A.: A general approach to fuzzy concepts. Math. Log. Quart. 50(3), 265–280 (2004)
17. Zhang, W.X., Ma, J.M., Fan, S.Q.: Variable threshold concept lattices. Inform. Sci. 177(22), 4883–4892 (2007)

Chapter 10
A Multi-attribute and Logic-Based Framework of Ontology Alignment

Marcin Pietranik and Ngoc Thanh Nguyen

Abstract. Ontology alignment is an issue that focuses on designating the way of migrating the contents of two ontologies. These structures can be treated as a method of decomposing some domain of interest and expressing its complexity by describing semantic correlations between extracted elements. The necessity of transforming information stored in two separated, independent ontologies arises when two computer systems incorporating such ontologies need to be integrated. Careful research on previously proposed methods made us think about a novel approach to this task, based on deeper analysis of basic building blocks of concepts, which are their attributes. This chapter is a comprehensive description of our ideas, conceptions and algorithms. We also give a brief commentary on preliminary results that will illustrate our contribution to the topic.

10.1 Introduction

In this chapter we want to give a comprehensive description of our method of aligning ontologies. Our solution is built around assigning formal semantics to concepts' attributes and analyzing relationship that can eventually occur between them. These relationships describe how attributes interact with each other and moreover, how much information that they express can be transformed.

The need of aligning ontologies arises when two computer systems need to be integrated or when they exchange content of their knowledge bases. Let's assume that there are two infrastructures with knowledge bases KB_1 and KB_2, both incorporating ontologies (respectively *O* and *O'*). When some user sends a request to KB_2, he expects that the answer will be returned in the format imposed by ontology *O'*. The

Marcin Pietranik · Ngoc Thanh Nguyen
Institute of Informatics, Wroclaw University of Technology, Wybrzeze Wyspianskiego 27, 50-370, Wroclaw, Poland
e-mail: marcin.pietranik@pwr.wroc.pl,
ngoc-thanh.nguyen@pwr.wroc.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 99–108.
springerlink.com

fact that necessary data may not be present within utilized system, but in the other one, must be completely transparent. Therefore, these cooperating systems need to exchange their contents and the second one must be able to find those parts of *O* that most accurately match to specific parts of *O'*.

From the formal point of view the problem of designating alignments between ontologies can be formulated as follows: *Having two ontologies O_1 and O_2, designating alignment between them consists in determining the set of tuples of the form $< c, c', M_C(c, c') >$, where c and c' are concepts from respective ontologies O_1 and O_2 and the real value $M_C(c, c')$ describes the degree to which concept c can be aligned to concept c'.*

In our approach we introduce the novel idea of assigning explicit semantics to attributes. In previous works that have been done in this field, authors concentrate mainly on developing variety of methods, which calculate similarities between concepts, based largely on their labels and relations that occur between them. Such approach has two major downsides. First, similarity from mathematical point of view is strictly symmetric, so prepared solutions also return symmetric mappings as a result. In our opinion it is not consistent with intuitive way of thinking about aligning ontologies - obviously there are situations in which it is much easier to find valid mappings from one ontology to another, than in the second direction. Secondly, all of the analyzed methods rely on one specific ontology representation, which is broadly accepted OWL language. This fact has some repercussions, among others, the impossibility of aligning ontologies that are not stored as OWL documents and the inability of utilizing such properties of ontologies that cannot be expressed in this format. We consider storing ontologies in flat text files, which in essence OWL files are, as not flexible enough. Such solution do not reflect the dynamics of the real world, that in our opinion is required when acting as a scheme that is expected to express the complexity of any domain. We claim that analyzing assignments of formal semantics to attributes is more expressive and open-ended than designating symmetric mappings based on variety of similarities that can be calculated. Hence, we propose a new formal format of expressing ontologies that overcomes restrictions of OWL standard. What is more, we also provide its implementation, utilizing modern web technologies. In the following chapter at first we will carefully define our approach to ontologies and than we will give detailed, formal description of every step we take to align them.

The main contribution of this chapter (that strictly follows our ideas presented in [9] and [10]) is identification of difficulties in comparing our ideas with previous approaches to ontology alignment. Throughout part 10.2 we present basic notions and definitions that act as a foundation for further sections. Part 10.3 is a description of our approach. In Part 10.5 we concentrate on describing problems that appear while attempting to compare our ideas with former solutions to considered topic. Short commentary about experimental environment that has been prepared is also given. In the last part we give a short summary and compact overview of upcoming work.

10.2 Basic Notions

Taking work that has been done in [6] and [11] as a starting point we define ontology as following:

$$O = (C, R, I) \tag{10.1}$$

where C is the set of concepts, R is a set of relations between them (defined as $R \subseteq C \times C$) and I is a set of instances. Following [12] the concept c from set C is defined as a triple:

$$c = (Id^c, A^c, V^c) \tag{10.2}$$

in which Id^c is a concept's label (an informal name of a concept), A^c is a set of it's attributes and V^c is a set of domains of attributes from A^c. The triple is further c called *concept's structure*.

Let's assume that there exists a finite set A of attributes and set V of their valuations. In further parts of this article a pair *(A,V)* will be denoted as *Real World* and every ontology O such that $\forall_{c \in C} A^c \subseteq A$ will be called *(A,V)-based*.

We assume that there exists a fixed set S of basic descriptions of attributes' semantics. A single item from this set is some, indivisible description given in natural language. We denote L_s as the formal language, using elements of S and logic operators $\neg, \vee, \wedge$. We can say that L_s is a sublanguage of the sentence calculus language.

Definition 1. By semantics of attributes within concepts we call a partial function:

$$S_A : A \times C \to L_s \tag{10.3}$$

Logic sentences are assigned to attributes when they are included in some concept's structure. Such definition allows us to reflect the variety of ways in which particular attribute can behave. For example, the attribute *Address* can act differently when occurs in the concept *Webpage* and concept *School*. Moreover, using these descriptions we are able to formally identify how attributes are related with each other. As in [9] we assume that we have two ontologies *O* and *O'* with respective sets of concepts *C* and *C'*. Below we present criteria that we use to distinguish possible associations.

Definition 2. Two attributes $a \in A^c, b \in A^{c'}$ are equivalent referring to their semantics (*semantical equivalence*) if the formula $S_A(a,c) \Leftrightarrow S_A(b,c')$ is a tautology for any two $c \in C, c' \in C'$. To mark this relation we will use the symbol $\equiv$.

For example, the attribute *Id* is equivalent to attribute *IdentificationNumber* and we can denote this fact as $LastName \equiv Surname$.

Definition 3. The attribute $b \in A^{c'}$ in concept $c' \in C'$ is more general than attribute $a \in A^c$ in concept $c \in C$ referring to their semantics (*semantical generality*) if the formula $S_A(a,c) \Rightarrow S_A(b,c')$ is a tautology for any two $c \in C, c' \in C'$. To mark this relation we will use the symbol $\uparrow$.

For example, the attribute *Surname* is more general than attribute *FullName*, because knowing someones' full name, we can easily get his surname but knowing his last name, we cannot designate both his first and last name. We denote this fact as $FullName \uparrow Surname$.

Definition 4. Two attributes $a \in A^c, b \in A^{c'}$ are contradictory referring to their semantics (*semantical contradiction*) if the formula $\neg(S_A(a,c) \wedge S_A(b,c'))$ is a tautology for any two $c \in C, c' \in C'$. To mark this relation we will use the symbol ↓.

For example, two attributes *Available* and *NotInStock* are in contradiction and we denote this fact as $IsActive \downarrow IsSuspended$.

10.3 Attribute-Based Concept Alignment

In this section we will present how we incorporate notions described throughout part 10.2 in the process of answering the question about the degree to which we can align concepts taken from two heterogeneous *(A,V)*-based ontologies. The core of our approach is the chain of functions $M_A^{c,c'}, M_A^c$ and M_C that calculate respectively the degree of alignment of two attributes, the degree of alignment for selected attribute from source ontology and the degree of alignment of two concepts. Having that in mind we have formulated a set of postulates that the function $M_A^{c,c'}$ must satisfy: *(i) The function* $M_A^{c,c'}$ *must not be symmetrical. (ii) If two attributes* a *and* b *are equivalent then* $M_A^{c,c'}(a,b) = M_A^{c,c'}(b,a) = 1$. *(iii) If* $a \uparrow b$ *and not* $a \equiv b$ *then* $M_A^{c,c'}(a,b) = 1$ *and* $M_A^{c,c'}(b,a) < 1$. These postulates are straightforward formalization of the intuitive way of thinking about transforming two representations of knowledge between each other. There are situations in which it is much easier to migrate data from particular source into another than in the other direction. The second postulate specifies such case taking into account the semantic relationships between attributes when one attribute is more general than another we are able to unequivocally designate its value. For example, owing someones date of birth we can always conveniently calculate his age, but it is impossible to designate date of birth more accurately than to year of birth owing only someone's age. Taking into consideration given postulates we have defined the function $M_A^{c,c'}$ as follows:

$$M_A^{c,c'}(a,b) = \begin{cases} 1 & \text{if } a \equiv b \\ 1 & \text{if } a \uparrow b \text{ and not } a \equiv b \\ 1 - d_S(S_A(a,c), S_A(b,c')) & \text{otherwise} \end{cases} \tag{10.4}$$

The value of the distance between two semantics (d_S) is calculated according to the algorithm firstly mentioned in our previous publication [12]. For detailed description please refer to [10].

The second step of our method is designate "the best match" for selected attribute from source concept within the set of attributes of the target concept. This is realized by the function $M_A^c : A^c \to [0,1]$ defined below:

$$M_A^c(a) = \begin{cases} \frac{1}{|Z^*|}\sum_{(a,b)\in Z^*} M_A^{c,c'} & \text{if } |Z^*| > 0 \\ M_A^{c,c'} & \text{if } |Z^*| = 0\text{, for } b = argmax_{b\in A^{c'}} M_A^{c,c'}(a,b) \wedge M_A^{c,c'}(a,b) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (10.5)$$

In which the working set Z^* is defined as follows:

$$Z^* = \{(a,b) : a \in A^c, b \in A^{c'}, b = argmax_{b\in A^{c'}} M_A^{c,c'}(a,b) \wedge M_A^{c,c'}(a,b) \geq T\} \quad (10.6)$$

Having in mind that the main goal of our work is to give a method for designating the degree to which particular concept from source ontology can be aligned to some other concept from target ontology (in other words- how much information can be reliably transformed) we have formulated two postulates that the function M_C must satisfy: *(i) The function M_C must not be symmetrical. (ii) Assuming the existence of concepts* c *and* c' *such that $A^c = a$ and $A^{c'} = b$ where $a \uparrow b$ and not $a \equiv b$ then the condition $M_C(c,c') \geq M_C(c'c)$ must be met.* By incorporating the two previous functions $M_A^{c,c'}$ and M_A^c and the listed postulates we were able to formulate the algorithm that is used to calculate the function M_C. Due to the limited space of the following chapter we will not provide the whole listing, but we will give an overview of key elements. The algorithm takes as input two sets of attributes (A^c and $A^{c'}$) and at first it removes the unnecessary redundancy from the source concept's structure A^c (for example - it discards attributes that are equivalent according to definition 2), creating auxiliary set $\overline{A_c}$. Then it returns the end result specified by the equation $M_C = \frac{\sum_{a\in\overline{A_c}} M_A^c(a)}{|\overline{A_c}|}$. For details please refer to [10].

10.4 Related Works

In [7] a classification of so-called ontology conflicts is given. These incompatibilities are the major reason of the difficulty while integrating two or more ontologies. This issue can be treated as a foundation for further works, because it allows us to identify with what situation the method of providing interoperability between ontologies will need to handle.

Variety of different methods of overcoming ontological diversity can be found in literature. The comprehensive survey of basic approaches can be found in [2]. A wide range of solutions have been proposed - from simple string comparison, structure analysis or more complex techniques involving aggregation of different aspects of ontology analysis. Recently, methods concerning ontology alignment have been gathered under OAEI initiative which provide a consistent way of evaluating approaches to ontology alignment stored in OWL format. Throughout subsequent years obtained results have been presented in dedicated proceedings available online ([13]). In publications referring to the topic of matching database schemas we can find issues related to attribute matching. Similarly to our ideas about identifying relationships between attributes in [4] authors approach to this issue with tokenization and lemmatisation of their labels and then utilization of external thesaurus, such as WordNet along with formal analysis of the taxonomy in which attributes and concepts are organized. An efficient algorithm that designates attribute correspondences is also provided. Another compelling method of analyzing the meaning of attributes has been presented in [5]. Authors formulate conditions for distinguishing relationships between attributes by analyzing their domains and valuations. They also provide a few examples and the definition of *uncertain semantic relationship* which is the extension of their basic idea.

Due to limited space we cannot provide more detailed description of results presented in former articles. For broader overviews of the work that has been done in this field please refer to our previous publications [9], [10], [11] and [12].

10.5 Evaluation Methodology

Former approaches to ontology alignment to evaluate and eventually rate obtained results mainly incorporated test data prepared by Ontology Alignment Evaluation Initiative ([13]). Available benchmarks are built on top of the idea of providing a set of large-scale ontologies and manually prepared, reference mappings. Ontologies included in this set are heavily differentiated - the inconsistencies between them spread across different naming conventions, varying hierarchies, mismatching instances, etc. Tested methods confront their results with expected alignments and then standard *Precision* and *Recall* measures are used to unequivocally compare them. This indicators are defined respectively as $Pr = \frac{|Ref \cap Al|}{|Al|}$ and $Re = \frac{|Ref \cap Al|}{|Ref|}$, where *Ref* is a set containing reference mappings that include pairs of concepts from two ontologies and *Al* is mappings designated by particular method that is evaluated.

The biggest downside of our method is its lack of possibility to strictly compare it with previous solutions in terms of aforementioned measures of *Precision* and *Recall*. We have identified few reasons of this situation. First of all, our method is based on semantics of attributes and former approaches did not incorporate this kind of information within their workflows. What is more, OWL standard (in which ontologies are typically stored) do not provide any formal mechanism to express such data (semantics of attributes given as explicit logic formulas). In [3] the comprehensive commentary of problems related to this standard has been given. Authors present many of the specific difficulties that occur while expressing ontologies using this format, among other syntactic diversity, expressivity limitations or naming ambiguity.

Previous approaches to ontology alignment concentrated mainly on the concept level - analyzing features directly related to them, such as: names, labels, descriptions, relations that connect them with each other, taxonomy in which they are organized, etc. What is more, all of the methods were based on data available in OWL standard, and do not attempt to extend it. Our solution focuses on deeper analysis of this level and moves the attention to processing attributes of concepts and the impact they can have on designating ontology alignment.

These are the reasons why we have prepared our novel approach to storing and handling ontologies based on solid, theoretical foundation. Obviously, to provide applicability of our ideas we have then implemented an experimental environment that allows convenient managing ontologies. This includes their creation, flexible concept handling or custom logic inference engine used for assigning semantics to attributes. Nevertheless, to reliably compare any two or more methods we must use solid and robust test data, that can be utilized as the input for algorithms, which need to be verified. The comparison is reliable and solid only if the same benchmarks is incorporated. In terms of testing ontology alignment methods, such testbed must contain ontologies with adequate number of concepts and provide plausibly prepared reference alignments. In order to prove our method with different input data than previously utilized, we can naturally prepare some number of ontologies by hand. Such approach could provide both data required by previous methods and semantic descriptions of attributes needed by our algorithms. Nevertheless such solution would be very time and resource consuming and what is more, it would always cause doubts in terms of its reliability and cardinality. In parallel, preparing manually such testbed, with hundreds of ontologies and concepts within them, is actually impossible.

Bearing in mind the necessity of illustrating somehow the correctness of our ideas, we have prepared a limited benchmark. It contains a set of concepts, each described by complementary set of its attributes. These attributes include a label (expressed in natural language, with no scrambling or language differences) and manually assigned semantics. Labels of concepts are based on OAEI benchmark ([13]). Then the simple method with which we want to compare our approach has been implemented. It incorporates string similarity and WordNet - a pair of tools frequently used by variety of aligning systems. The overall degree of aligning two concepts is given by the equation below:

$$M_{simple}(c,c') = \frac{1}{2|A^c|}\sum_{a\in A^c}\left(\max_{b\in A^{c'}} sim_s(a,b) + \max_{b\in A^{c'}} sim_w(a,b)\right) \qquad (10.7)$$

The idea is straightforward - the algorithm crawls through a set of attributes of concept from source ontology and selects attributes of concept from target ontology for which two given similarities (sim_s is Levenshtein's string similarity and sim_w is WordNet acquired value of path length similarity) are maximal. The preparatory phase for analyzing labels include extracting lemmas and removing stop words (such as prepositions). Comparative results are presented in tables 10.1 and 10.2.

These results have been collected using aforementioned experimental environment. During the development process we have decided to incorporate as a persistency layer the document-oriented database MongoDB. Unlike in relational systems that keep their data in unified tables, assumed tool do not require any kind of schema. Loosely coupled documents, that are somehow similar, are grouped in collections, which act as aggregators for information of the same type. Beside the fact that they are included in one category, they do not need to share any inner description (for example - elements of business cards' collection may include variety of content). In our opinion, lack of schema and open-ended structure of MongoDB ideally corresponds to requirements for storing ontologies. On top of raw database we have built a wrapper that allows processing semantics and executing all of the algorithms from

Table 10.1 Comparative results (1)

		Target concepts				
		Academic	Article	Collection	Conference	LectureNotes
Source concepts	hazdn	0.5 (0.269)	0.957 (0.639)	0.76 (0.402)	0.729 (0.44)	0.8 (0.485)
	scds	0.5 (0.22)	0.722 (0.412)	0.641 (0.3678)	0.8 (0.253)	0.966 (0.554)
	sqxsqkd	0.5 (0.278)	0.848 (0.384)	0.9 (0.567)	0.608 (0.344)	0.683 (0.45)
	TechnicalReport	0.5 (0.252)	0.843 (0.635)	0.625 (0.434)	0.625 (0.285)	0.625 (0.44)
	xsqlknk	0.5 (0.246)	0.875 (0.85)	0.675 (0.402)	0.6 (0.407)	0.6 (0.405)
	zadazxn	0.59 (0.434)	0.583 (0.317)	0.5 (0.239)	0.708 (0.549)	0.555 (0.274)

Table 10.2 Comparative results (2)

		Target concepts			
		MasterThesis	Publication	School	TechReport
Source concepts	hazdn	0.7 (0.396)	0.8 (0.37)	0.5 (0.295)	0.7 (0.403)
	scds	0.6 (0.31)	0.583 (0.293)	0.536 (0.293)	0.6 (0.228)
	sqxsqkd	0.8 (0.325)	0.7 (0.284)	0.5 (0.321)	0.7 (0.311)
	TechnicalReport	0.75 (0.426)	0.719 (0.61)	0.5 (0.266)	0.968 (0.432)
	xsqlknk	1 (0.705)	0.675 (0.547)	0.6 (0.283)	0.7 (0.409)
	zadazxn	0.583 (0.412)	0.5 (0.234)	0.95 (0.623)	0.5 (0.417)

section 10.3. The last element is the interface available both as the web application and REST-based API. It is much easier to interact with our system and therefore to utilize the method in variety of applications such as integrating federated data-warehouses or any collective intelligence problems [8].

We have used the slightly modified input data described in [10]. We have also swapped source and target ontologies with each other. The first number in cells from tables 10.1 and 10.2 is the result of our method. In brackets we give the score obtained by similarity measure from equation 10.7. As easily seen, gathered results prove correctness of our method. They show that we are able to state more unequivocally which concept from source ontology can be matched to concepts from target ontology. The output of our algorithm is more unambiguous. It is caused by the fact, that our method does not rely on labels of attributes or any other labels. Therefore, it can find good matches not only when two labels are similar (e.g. *DateOfBirth* and *BirthDate*). Obviously, there are situation in which previous methods (described for example in [13] or [2]) give better results, but bear in mind that our solution analyzes only single concepts and not whole ontologies. What is worth emphasizing, it is not any kind of similarity - in contradiction to assumptions from [2] we do not provide symmetrical results. Our aim is different - we want to answer the question about the amount of information that can be reliably transformed. That is the reason why sometimes our solution return high values for few different concepts. This fact does not imply that these concepts are similar, but that data stored within them can be migrated. For example - imagine high value of alignment degree for concepts *Article* and *University* and the situation in which the end user request the information about articles which scope is the main interest of scientific workers of some university. In such situation our method appears to be better suited to handle it.

Interesting situation also appear while trying to align some class from source ontology to concept *Academic* (Table 10.1). As stated - we describe attributes by assigning formal semantics that are eventually transformed into conjunctive normal form (that can be treated as sets of clauses). Each clause can be decomposed into sets of positive and negative literals. The eventual distance is the average distance between these sets. Multiple *0.5* values are caused by lack of negative literals within semantics of processed attributes - such situation causes processing two empty sets (that are obviously identical) which increases the overall value. For more details of used test data we encourage to refer to [10].

Due to the fact that our method focuses on deeper analysis of concepts' structure we can become independent from analyzing ontology as a whole structure. Therefore, aligning two concepts do not require any kind of additional information about other concepts. The method takes into account only the knowledge that is available within structures describing selected concepts. We also become independent from any external knowledge base (such as WordNet) and we can overcome problems associated with handling labels of concepts that are compound words (these kinds of issues are closely related to natural language processing and they are frequently beyond the scope of ontology alignment). What becomes clear thanks to presented results - these structures provide powerful, expressive and complex semantic description that is sufficient to designate proper mappings.

10.6 Summary and Future Works

In this chapter we have presented our novel approach to ontology alignment. We base our ideas on assigning explicit, formal semantics to attributes, which are basic elements of concepts within different ontologies. The main contribution is careful analysis of experimental issues that occur while testing our method. We have provided a set of difficulties that prevent us from strict comparing our solution with previously developed algorithms. Moreover, we have given a set of initial ideas about how we want to verify described approach.

In the future we want to concentrate on gathering more comprehensive set of input data. We want to incorporate information taken from our other research interests such us integrating schemas of federated data warehouses. Simultaneously, we will continue to develop our experimental environment, extending it with possibility to gather such data automatically. On the formal side, we want to treat our approach as a foundation of any issue related to ontology alignment. The easiest task to identify is proposing a method of aligning relations between concepts. Concurrently, we will work on reducing the complexity of finding matches between concepts. Taking [1] as a starting point, the base idea is to divide sets of concepts from two ontologies into disjoint groups and search for valid mappings only between entities that have been classified into the same category.

Acknowledgements. This research project was supported by Polish Ministry of Science and Higher Education under grant no. N N519 407437 (2009-2012) and by grant No. 2011/01/N/ST6/01438 from the National Science Centre, Poland.

References

1. Duong, T.H., Jo, G.S., Jung, J.J., Nguyen, N.T.: Complexity analysis of ontology integration methodologies: A comparative study. Journal of Universal Computer Science 15, 877–897 (2009)
2. Euzenat, J., Shvaiko, P.: Ontology Matching, 1st edn. Springer, Heidelberg (2007)
3. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: OWL 2: The next step for OWL. Web Semantics: Science, Services and Agents on the World Wide Web 6, 309–322 (2008)
4. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: Semantic Schema Matching. In: Meersman, R. (ed.) OTM 2005. LNCS, vol. 3760, pp. 347–365. Springer, Heidelberg (2005)
5. Magnani, M., Rizopoulos, N., Mc.Brien, P., Cucci, F.: Schema Integration Based on Uncertain Semantic Mappings. In: Delcambre, L.M.L., Kop, C., Mayr, H.C., Mylopoulos, J., Pastor, Ó. (eds.) ER 2005. LNCS, vol. 3716, pp. 31–46. Springer, Heidelberg (2005)
6. Nguyen, N.T.: Advanced Methods for Inconsistent Knowledge Management (Advanced Information and Knowledge Processing). Springer (2008)
7. Nguyen, N.T.: Conflicts of Ontologies – Classification and Consensus-Based Methods for Resolving. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 267–274. Springer, Heidelberg (2006)
8. Nguyen, N.T.: Processing Inconsistency of Knowledge in Determining Knowledge of a Collective. Cybernetics and Systems 40(8), 670–688 (2009)
9. Pietranik, M., Nguyen, N.T.: A Distance Function for Ontology Concepts Using Extension of Attributes' Semantics. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 623–632. Springer, Heidelberg (2011)
10. Pietranik, M., Nguyen, N.T.: A Method for Ontology Alignment Based Attribute Semantics. Cybernetics and Systems (to appear, 2012)
11. Pietranik, M., Nguyen, N.T.: Attribute Mapping as a Foundation of Ontology Alignment. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS(LNAI), vol. 6591, pp. 455–465. Springer, Heidelberg (2011)
12. Pietranik, M., Nguyen, N.T.: Semantic Distance Measure between Ontology Concept's Attributes. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part I. LNCS(LNAI), vol. 6881, pp. 210–219. Springer, Heidelberg (2011)
13. Shvaiko, P., Euzenat, J., Giunchiglia, F., Stuckenschmidt, H., Mao, M., Cruz, I.: Proceedings of the 5th International Workshop on Ontology Matching (2010), http://ceur-ws.org/Vol-689/

Chapter 11
Application of an Ontology-Based Model to a Wide-Class Fraudulent Disbursement Economic Crimes

Czesław Jędrzejek and Jarosław Bąk

Abstract. We analyze the whole transaction cycle starting from an agreement between two companies or a company and a person up to a physical process of executing a payment. Previously, we considered only fraudulent disbursement economic crime perpetrated by management. In this work we generalize our minimal ontology model of economic crimes to dealing with crimes committed by employees not in managerial positions. Extending the concepts and relations of the minimal model and using appropriate rules, we are able to map crime activity options (roles of a particular type of employees). This makes it possible to reason penal code sanctions using activities and roles of persons in a crime. Although the work pertains mostly to the Polish Penal Code, the model is able to distinguish between embezzlement and larceny (and thus in some aspect encompasses asset misappropriation not only money misappropriation). Prospects on future reasoning capabilities of the tool will be presented.

11.1 Introduction

There is much progress in studying economic crime mechanism in order to prevent, detect and prosecute them. According to ACFE [1] economic crimes cost typical organization 5% of its annual revenue.

As far as typology there exist many classifications. ACFE in its Occupational Fraud and Abuse Classification System divides crimes into 3 categories: corruption, asset misappropriation, fraudulent statements. Next, asset misappropriation is divided into cash and non-cash (in our terminology money, the cash being form of money and non-monetary assets). The large category of money related crimes are fraudulent disbursements.

Czesław Jędrzejek · Jarosław Bąk
Poznań University of Technology,
Institute of Control and Information Engineering,
Poznań, Poland
e-mail: {Czeslaw.Jedrzejek,Jaroslaw.Bak}@put.poznan.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 109–118.
springerlink.com

According to PricewaterhouseCoopers [2] economic crimes can be divided into: asset misappropriation, accounting fraud, bribery and corruption, intellectual property infringement, money laundering, tax fraud, illegal insider trading, market fraud involving cartels colluding to fix prices, espionage, and other. As we can see the classification of PwC is wider than ACFE.

In general mapping technical economic crime schemes into legal meaning is not straightforward. ACFE classification does not have explicit embezzlement category since several schemes fit this crime. In legal acts and the FBI classification there are broad categories of a corporate crime, and a white collar crime or narrower categories (for the most recent official version of US Code made available by the US House Representatives).

Modeling economic crimes is difficult, however, increasingly better statistics and analysis of schemes provides systematic material which makes this task easier to model and code into an expert system.

In particular, in [3-5] flowcharts of activities for important classes of crimes had been developed. Because, fraudsters use many types of schemes, techniques and transactions to achieve their goals, it has seemed impossible to construct a simple conceptual model of any generality. Only recently has the integrated use of semantics expressed by means of ontologies and rules achieved the capability of analyzing large practical problems, such as applying reasoning over legal sanctions on the basis of investigation facts and rules appearing in penal codes.

In a series of works, using SROIQ logic we constructed the so-called minimal model of very frequent and important classes of the following crimes, using which one can reason up to the sanctions. These are:

a) fraudulent disbursement economic crime, perpetrated by management [6]
b) fraudulent disbursement economic crime, perpetrated by management and money laundering [7].

In the previous works [6-7] we basically used concepts best suited to map activities to crimes sanctioned by the Polish Penal Code [8]. Thus the general economic crime was "causing damage to a company" (Polish: działanie na szkodę spółki) in the form of asset misappropriations (Polish: sprzeniewierzenie majątku), including fraudulent disbursement (Polish: sprzeniewierzenie pieniędzy). Such crime is very widespread and intractable across countries and industries. In a 2009 survey [2], using economic crime as the top category, asset misappropriation constituted two-thirds of all economic crimes. Such a crime is often accompanied by money laundering schemes. According to ACFE [1] asset misappropriation constitute 86-89% (depending on the region) of Occupational Fraud classification.

The chapter is organized as follows. Section 11.2 presents main features of extension of our minimal model to fraudulent disbursement economic crime, perpetrated by management and non-management employees. Section 11.3 presents the concerns with the rules necessary to describe this l case (that contain all sensible mechanism options of this class of crime). Section 11.4 presents examples of questions and the reasoning system answers. Conclusions and future work are presented in Section 11.5.

11.2 The Model

In this work we extend our previous ontology and the model [6-7] to fraudulent disbursement economic crime, perpetrated by management and non-management employees. At this time we skip schemes related to cash (bills and coins) and checks, and concentrate on invoice related fraud (basically encompassing: billing schemes, and shell company schemes) [1], [3-5]. The fraudulent disbursement schemes which were currently added to our model are presented on Fig. 11.1. In addition to the ontology of the model described in [6-7] the following classifications are added:

Positions in companies

These are:

- managerial, non-managerial financial, non-managerial administrative,
- employed in the department transaction is taking place or employed in the department transaction is not taking place.

A company consists of one or more departments or other administrative units (for example: construction sites). Transaction may be related to: goods, services.

We assume the following workflow of transaction cycle activities:

- Prior to issuing an invoice:
  1. Involved Department plans to buy goods or subcontract a service.
  2. Director accepts the order (above a certain threshold a Member of the Board accepts the order (accompanied by agreement).
  3. Financial clerk accepts the order (if authorized by a Chief Financial Officer), otherwise the authorization is done by Chief Financial Officer.
  4. Order goes into company books.
- After accomplishing the contract (order):
  1. Registration of incoming document in the registrar's office (for example: invoice).
  2. Registration of invoice in the financial department.
  3. Transferring invoice to department responsible for a given purchase (alternative route invoice may be brought directly to department responsible for a given purchase).
  4. Generalized Acceptance of payment:
     a) Verification that the goods have been delivered according to the order or statement that service/work has been performed (statement of construction work acceptance).
     b) Acceptance of payment of manager/board: multistage.
     c) By director, member of the executive body, or Chief Financial Officer.
  5. Transferring signed invoice to unit responsible for paying bills (making money transfers).
  6. Making money transfers by an authorized administrative officer.

Fraudulent disbursement in transaction cycle

How this normal cycle can be broken without participation of managers:

- administrative officer whose is authorized to make payment makes payment for real goods/service to accomplice or phony company in which he/she has control (or in conspiracy such as Hydra and Hermes in [7]) instead to legitimate payee (with or without forged accepted invoice – crime by administrative officer whose is authorized to make payment physically).
- administrative officer whose is authorized to make a payment makes a payment on a forged by someone acceptance document.
- someone in a company steals passwords to the company account and makes a payment based on a forged acceptance of payment (alone or with an accomplice that forged acceptance of payment) – this option exists in Figure 11.1, but it has not yet been included in rules.

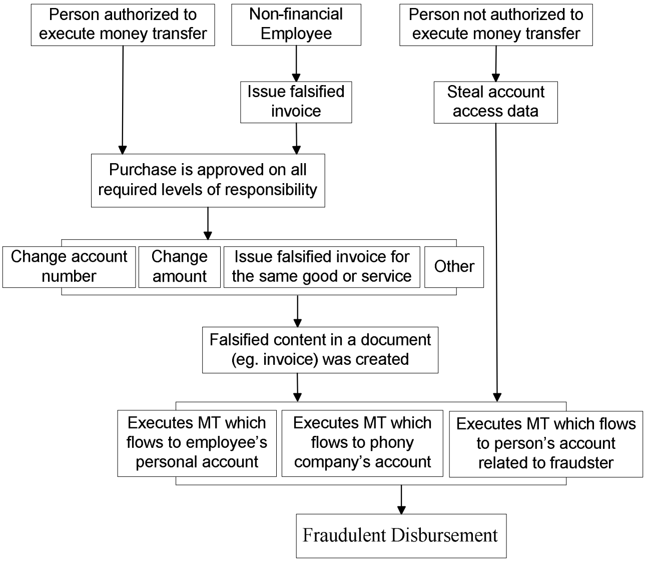

Person authorized to execute money transfer
Non-financial Employee
Person not authorized to execute money transfer
Issue falsified invoice
Steal account access data
Purchase is approved on all required levels of responsibility
Change account number
Change amount
Issue falsified invoice for the same good or service
Other
Falsified content in a document (eg. invoice) was created
Executes MT which flows to employee's personal account
Executes MT which flows to phony company's account
Executes MT which flows to person's account related to fraudster
Fraudulent Disbursement

Fig. 11.1 Some of Fraudulent Disbursement schemes in our minimal model ontology.

The flow of activities presented in Figure 11.1 could by compared with equivalent "False Billings from Shell Companies" (Figure 12.1) and "Invoice Purchasing Scheme" (Figure 12.3) of [5]. On one hand, it allows for handling fewer micro schemes compared to [5], partly because we do not consider cash and check schemes, but on the other hand, because of much richer semantics, it allows for mapping activities into legal sanctions. Both approaches disregard tax issues [9]. In most countries tax penalties differ between false and fictitious invoices, which have to be properly defined. In our approach a document has a broad meaning – could be a digital document or a record with legally meaning data. Compared to ontology accessible in [10] we introduce the following new concepts:

- PersonAuthorizedToExecuteMT
- BankAccount
- FalsifiedBankAccount
- FinancialClerk
- Chief FinancialOfficer – may be at the level of a member of Executive Board or a Director
- PersonAuthorizedToAccountantActivity
- PersonNotAuthorizedToAccountantActivity
- FalsifiedMoneyTransfer
- FalsifiedContent (FalsifiedSignature, FalisfiedAccountNumber or others)
- FinancialEmployee
- NonFinancialEmployee

Relations:

- hasExecutedMT (MT, money transfer)
- isOwnerOf
- doesntWorkFor
- signedOnResponsibilityLevel
- hasDefinedAccount
- createdBy (pertains to signatures, documents, records in books, etc.).

Limited implementation of our model in OWL ontology with SWRL rules is available at RuleML Challenge 2011 Demo [10] site: http://draco.kari.put.poznan.pl.

11.3 Set of Rules

Rules pertain to various options of paying invoice to accounts related to fraudsters, who are authorized to make transfers. Letters preceded by a question mark represents variables.

1. This rule concerns a real invoice but the payment goes to account belonging to a fraudster instead of company's account.

```
Invoice(?i), PersonAuthorizedToExecuteMT(?p), isSignedBy(?i, ?p), Transaction(?t), hasInvoive(?t, ?i), hasMoneyTransfer(?t, ?mt), transactionFrom(?t, ?c1), transactionTo(?t, ?c2), MoneyTransfer(?mt), hasExecutedMT(?p, ?mt), worksFor(?p, ?c1), Company(?c1), flowsFrom(?mt, ?a1), BankAccount(?a1), isOwnerOf(?c1, ?a1), flowsTo(?mt, ?a2), BankAccount(?a2), isOwnerOf(?p, ?a2), DifferentFrom(?c1, ?c2) → FalsifiedBankAccount(?a2), FalsifiedMoneyTransfer(?mt)
```

2. This rule concerns a forged invoice but the payment goes to an account belonging to a fraudster. The person making the transfer is a forger.

```
Invoice(?i), PersonAuthorizedToExecuteMT(?p), hasForged(?p, ?i), MoneyTransfer(?mt), hasExecutedMT(?p, ?mt), worksFor(?p, ?c1), Company(?c1), flowsFrom(?mt, ?a1), BankAccount(?a1), isOwnerOf(?c1, ?a1), flowsTo(?mt, ?a2), BankAccount(?a2), isOwnerOf(?p, ?a2)
→ FalsifiedBankAccount(?a2), FalsifiedMoneyTransfer(?mt)
```

3. The rule indicates two transfers based on one real invoice: one to legitimate payee and one to an account belonging to a fraudster.

```
Invoice(?i), PersonAuthorizedToExecuteMT(?p), Transaction(?t1), hasIn-
voive(?t1, ?i), Transaction(?t2), hasInvoive(?t2, ?i), hasMoneyTransfer(?t1,
?mt1), hasMoneyTransfer(?t2, ?mt2), hasExecutedMT(?p, ?mt1), hasExecutedMT(?p,
?mt2), worksFor(?p, ?c1), Company(?c1), Company(?c2), transactionTo(?t1, ?c2),
flowsFrom(?mt, ?a1), BankAccount(?a1), isOwnerOf(?c1, ?a1), flowsTo(?mt1,
?a2), isOwnerOf(?p, ?c2)flowsFrom(?mt2, ?a1), BankAccount(?a2), isOwnerOf(?c1,
?a1), flowsTo(?mt2, ?a2), isOwnerOf(?p, ?a2), DifferentFrom(?c1, ?c2)
→ FalsifiedMoneyTransfer(?mt), FalsifiedBankAccount(?a2)
```

4. Assignment of the highest level of responsibility of approval of a given document. Depends on a hierarchy of responsibilities in a company. In result we obtain a level on which the document was signed.

```
Document(?d), Person(?p1), Person(?p2), isSignedBy(?d, ?p1), isSignedBy(?d,
?p2), hasLevelOfResponsibility(?p1, ?l1), hasLevelOfResponsibility(?p2, ?l2),
lessThan(?l1, ?l2), DifferentFrom(?p1, ?p2)
→ signedOnResponsibilityLevel(?d, ?l2)
```

5. Unauthorized person makes a transfer. The payment goes to account belonging to a fraudster.

```
PersonNotAuthorizedToAccountantActivity (?p), MoneyTransfer(?mt), hasExecu-
tedMT(?p, ?mt), worksFor(?p, ?c1), Company(?c1), flowsFrom(?mt, ?a1), BankAc-
count(?a1), isOwnerOf(?c1, ?a1), flowsTo(?mt, ?a2), BankAccount(?a2), isOwne-
rOf(?p, ?a2) → FalsifiedBankAccount(?a2), FalsifiedMoneyTransfer(?mt)
```

6. This rule concerns a real or forged invoice; but the payment goes to account belonging to a fraudster ?p1 or ?p3 (?p3 could be equivalent to ?p1).

```
PersonAuthorizedToAccountantActivity (?p1), MoneyTransfer(?mt), hasExecu-
tedMT(?p1, ?mt), worksFor(?p, ?c1), Company(?c1), Company(?c2), Invoice(?i),
isIssuedBy(?i, ?c2), isReceivedBy(?i, ?c1), DifferentFrom(?c1, ?c2), hasDefi-
nedAccount(?i, ?a2), isOwnerOf(?c2, ?a2), BankAccount(?a1), BankAccount(?a2),
BankAccount(?a3), isOwnerOf(?c1, ?a1), isOwnerOf(?c2, ?a2), flowsTo(?mt, ?a3),
isOwnerOf(?p3, ?a3), Person(?p3), DifferentFrom(?a1, ?a2), DifferentFrom(?a1,
?a3), DifferentFrom(?a2, ?a3)
→ FalsifiedBankAccount(?a3), FalsifiedMoneyTransfer(?mt)
```

7. Money transfer based on real or fictitious invoice but the payment goes to falsified bank account. In result we obtain a FalsifiedTransaction which joins a transaction and a money transfer.

```
Transaction(?t), hasMoneyTransfer(?t, ?mt), FalsifiedMoneyTransfer(?mt)
→ FalsifiedTransaction(?t)
```

8. The sanction according to the Polish Penal Code for a person authorized to execute money transfer, art. 284§2 (embezzlement).

```
FalsifiedMoneyTransfer(?mt), PersonAuthorizedToExecuteMT (?p), hasExecu-
tedMT(?p, ?mt), Art_284_2(?a), flowsTo(?mt, ?ac), BankAccount(?ac), isOwne-
rOf(?p, ?ac) → fallsUnder(?p, ?a)
```

9. The rule stating that if invoice was forged or falsified so is the money transfer (do not consider a case when invoice (complex document) was forged or falsified, money transfer not made, or not yet made). Person(?p1) and Person(?p2) could be the same or different (inComplicity, not yet defined in this rule).

```
ComplexInternalLegalDocument(?d), Person(?p1), Person(?p2), hasForged(?p1,
?d), MoneyTransfer(?mt), Transaction(?t), hasInvoice(?t, ?d), hasMoneyTrans-
fer(?t, ?mt), hasExecutedMT(?p2, ?mt)→ FalsifiedMoneyTransfer(?mt)
```

10. This rule states that if an Unauthorized person makes a transfer he/she is subjected to art. 270§1 of Polish PC (an aspect of larceny in US).

```
Art_270_1(?a), Docment(?d), PersonNotAuthorizedToAccountantActivity(?p), has-
Forged(?p, ?d) → fallsUnder(?p, ?a)
```

11. The rule defining FalsifiedComplexInternalLegalDocument – if one of signatures is forged the document is forged.

```
ComplexInternalLegalDocument(?d), FalsifiedContent(?f), contains(?d, ?f)
→ FalsifiedComplexInternalLegalDocument(?d)
```

12. The rule defining a FalsifiedComplexInternalLegalDocument – if ComplexInternalLegalDocument was forged then it is falsified.

```
ComplexInternalLegalDocument(?d), Person(?p), hasForged(?p, ?d)
→ FalsifiedComplexInternalLegalDocument(?d)
```

13. Rule defining that a person falsified document/record.

```
Person(?p), FalsifiedContent(?f), Document(?d), contains( ?d, ?f), created-
By(?f, ?p) → hasForged(?p, ?d)
```

14. The sanction art. 284§1 for making a falsified money transfer to an accomplice.

```
FalsifiedBankAccount(?a), FalsifiedMoneyTransfer(?mt), Art_299_1(?art2),
Art_284_1(?art2), Person(?p2), PersonAuthorizedToAccountantActivity(?p1), ha-
sExecutedMT(?p1, ?mt), isOwnerOf(?p2, ?a), isRelatedTo(?p1, ?p2), Diffe-
rentFrom(?p1, ?p2) → fallsUnder(?p1, ?art1), fallsUnder(?p2, ?art2)
```

15. The sanction – art. 271§1 for falsifying a document.

```
Art_271_1(?a), Document(?d), Person(?p), isIssuedBy(?d, ?p), contains(?d, ?f),
FalsifiedContent(?f), createdBy(?f, ?p) → fallsUnder(?p, ?a)
```

16. The sanction for a person who falsified a document (invoice) art. 270§1 but this document was accepted by a superior.

```
NonFinancialEmployee(?e), Document(?d), hasForged(?e, ?d), signedOnResponsibi-
lityLevel(?d, ?l1), signedOnResponsibilityLevel(?d, ?l2), hasLevelOfResponsi-
bility(?e, ?l1), lessThan(?l1, ?l2), Art_270_1(?a) → fallsUnder(?e, ?a)
```

17. The sanction for a person who falsified a document (invoice) but this document was accepted by a superior (who knew that a document was falsified). They are subjected to art. 270§1 and art. 271§1, respectively.

```
NonFinancialEmployee(?e), Document(?d), hasForged(?e, ?d), Art_270_1(?a1),
Art_271_1(?a2), signedOnResponsibilityLevel(?d, ?l1), FalsifiedContent(?f),
signedOnResponsibilityLevel(?d, ?l2), hasLevelOfResponsibility(?e, ?l1), less-
Than(?l1, ?l2), PersonAuthorizedToAccountantActivity(?p), hasLevelOfResponsi-
bility(?p, ?l2), knowsAbout(?p, ?f), createdBy(?f, ?e)
→ fallsUnder(?e, ?a), fallsUnder(?p, ?a2)
```

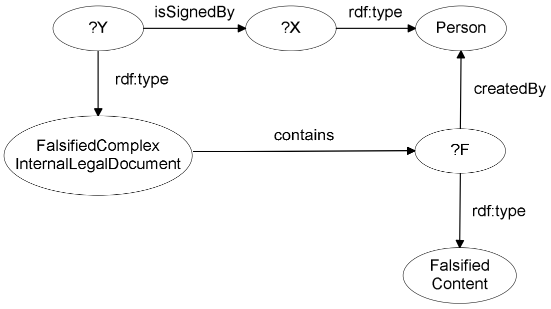


Fig. 11.2 A person (working no matter in what position) who falsified and signed a document/account (nature of which is not specified).

11.4 Queries

Based on the extended minimal model ontology (not yet public) and rules presented in Section 11.3 we are able to reason on various aspects of person's activities, status of transactions and legal sanctions. So far we have verified consistency of the ontology. We may at this stage propose various queries compatible with ontology. Here we present two examples of queries (Fig. 11.2 and 11.3). These questions allow for asking a position of a suspect person in the transaction workflow (transaction process cycle), and indicate possible sanctions.

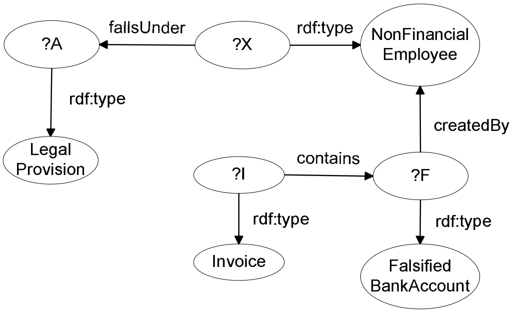


Fig. 11.3 A person (not working in financial department who made a transfer to illegitimate account) and thus is subject to sanctions (?A)

11.5 Conclusions

We extended the minimal economic crime minimal model to a wider typology. Currently, we are working on implementation of the knowledge base and generation of artificial data to test our approach. This would follow [7], where we successfully accomplished the reasoning functionality of the system.

In future works we will present and evaluate these results, as well as discuss scalability and discuss some metrics (the increase of ontology and number of rules with wider typology of crimes). The importance of this work is not in formal progress in reasoning but demonstrating that it is possible to build ontologies for nontrivial cases (that contain all sensible mechanism options of given classes of crimes). We are not aware of work of comparable complexity in legal area.

We hope that the knowledge base we develop will be amenable to extensions that would make it highly practical. This would, however, require harmonization of concepts between interested parties. For tax taxonomy, TaxML (not even ontology) this has not been feasible [11]. In order to have full practical importance, the system would have to go one level deeper, i.e. arrive at facts reasoned from available data (bookkeeping data, banking data, testimonies on relationships, etc.). In a much more limited model this has been done for fuel fraud [12]. It is important to ask why in law area we deal with such a detailed description is possible? The reason is that there is no unintentional fraud, and evidence we consider are documents and bank accounts, something that is traceable. If we have to consider details of intensions such as appearing in 18 definitions of money laundering in various legal systems collected in [13] the problem will became hard (as referred in [14]). Suppose we would like to distinguish between activities whose purpose is "Hide the proceeds" or "Make it appear legal". We cannot define these notions in general. They would depend on context, which would make us go into deep details, which would require rules of impossible to handle complexity (in an efficient way).

Acknowledgments. This work was supported by PUT 45-085/11 DS-PB grant.

References

[1] Association of Certified Fraud Examiners, Report to the Nations (2010), http://www.acfe.com/rttn/2010-rttn.asp

[2] PricewaterhouseCoopers' Global economic crime survey (2009), http://www.pwc.com/gx/en/economic-crime-survey/downloadeconomic-crime-people-culture-controls.jhtml

[3] Wells, J.T.: Billing Schemes. Journal of Accountancy, Parts 1- 4 194(1), 76–79, (2), 72–74, (3), 96–98, (4), 105–109 (respectively)

[4] Wells, J.T.: Corporate Fraud Handbook: Prevention and Detection, 2nd edn. Wiley, John & Sons (2007)

[5] Kranacher, M.-J., Riley, R., Wells, J.T.: Forensic Accounting and Fraud Examination. Wiley, John & Sons (2010)

[6] Bak, J., Jedrzejek, C.: Application of an Ontology-Based Model to a Selected Fraudulent Disbursement Economic Crime. In: Casanovas, P., Pagallo, U., Sartor, G., Ajani, G. (eds.) AICOL-II/JURIX 2009. LNCS(LNAI), vol. 6237, pp. 113–132. Springer, Heidelberg (2010)

[7] Bak, J., Jedrzejek, C., Falkowski, M.: Application of an Ontology-Based and Rule-Based Model to Selected Economic Crimes: Fraudulent Disbursement and Money Laundering. In: Dean, M., Hall, J., Rotolo, A., Tabet, S. (eds.) RuleML 2010. LNCS, vol. 6403, pp. 210–224. Springer, Heidelberg (2010)

[8] The Penal Code. Ustawa z dnia 6 czerwca 1997 r. Kodeks karny (1997) (in Polish)

[9] False and Fictitious Invoices IOTA Report for Tax Administrations, Budapest (2009), http://www.ujp.gov.mk/uploads//False%20and%20Fictitious%20Invoices.pdf

[10] Bak, J., Falkowski, M., Jedrzejek, C.: The SDL Library: Querying a Relational Database with an Ontology, Rules and the Jess Engine. In: Bragaglia, S., Damásio, C., Montali, M., Preece, A., Petrie, C., Proctor, M., Straccia, U. (eds.) Proceedings of the 5th International RuleML2011@BRF Challenge, co-located with the 5th International Rule Symposium, Fort Lauderdale, Florida, USA, November 3-5, vol. 799 (2011), http://ceur-ws.org/Vol-799/

[11] The TaxML Working group has been deactivated, before accomplishing a standard, http://www.oasisopen.org/committees/workgroup.php?wg_abbrev=taxtasc

[12] Jedrzejek, C., Falkowski, M., Smolenski, M.: Link Analysis of Fuel Laundering Scams and Implications of Results for Scheme Understanding and Prosecutor Strategy. In: 22nd International Conference on Legal Knowledge and Information Systems, JURIX 2009, Rotterdam, The Netherlands (2009)

[13] Unger, B., Busuioc, E.M.: The Scale and Impacts of Money Laundering. Edward Elgar (2007)

[14] Breuker, J.: Dreams, awakenings and paradoxes of ontologies, invited talk presentation. In: 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques (2009), http://ontobra.comp.ime.eb.br/apresentacoes/keynoteontobra-2009.ppt

Chapter 12
Modelling Failures of Distributed Systems in Stochastic Process Algebras*

Jerzy Brzeziński and Dariusz Dwornikowski

Abstract. The article presents a discussion on modelling failures of distributed systems in stochastic process algebras. The discussion is important from the point of view of proper representation of failures in order to express their impact on system's performance. The article presents formal definitions of failure models, as well as examples that are represented in one of the most popular stochastic process algebras — PEPA.

12.1 Introduction

One of the tasks while constructing distributed systems is ensuring that they will perform well. This can be done by means of simulations, tests of the working implementation of the system, or by using formal modelling. Stochastic process algebras (SPA) [1] are one of formal methods used for performance modelling of distributed systems. A modeller describes system's behaviour by means of states, actions and operators. Timing of actions is expressed by an exponential distribution parameter associated with every action. Structural operational semantics describe ways of transforming the model notation to a corresponding Labelled Transition System (LTS), which is then transformed to a Continuous Time Markov Chain (CTMC). CTMCs can be then solved by means of steady-state analysis, so interesting performance measures can be derived. However, for the steady-state solution to be possible, the CTMC has to be ergodic, i.e. every state has to be aperiodic and positive recurrent. It means that the model the CTMC was derived from, also needs to be

Jerzy Brzeziński · Dariusz Dwornikowski
Institute of Computing Science,
Poznań University of Technology,
Piotrowo 2, 60-965 Poznań, Poland
e-mail: {Jerzy.Brzezinski,Dariusz.Dwornikowski}@cs.put.poznan.pl

* This work has been partially supported by the Polish Ministry of Science and Higher Education within the European Regional Development Fund, Grant No. POIG.01.03.01-00-008/08 IT-SOA https://www.soa.edu.pl/

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 119–129.
springerlink.com

recurrent. Practically it means that the model cannot have any absorbing states that would cause a deadlock. In the case of fault-tolerant distributed systems modelling, fulfilling this requirement is challenging.

It is due to the fact that in distributed systems some failure types do show terminating behaviour, hence they cannot be directly modelled in SPAs that rely on CTMCs semantics. On the other hand, there exist failures that are strictly associated with basic distributed systems primitives, such as sending or receiving data. In that case the meaning of states and actions is important, not just their pure state-space representation. Stochastic process algebras do not deal with semantics, so expressing such failures is problematic, especially when automatic failure injection into models is considered.

In the context of these characteristics, failure modelling in SPAs is not trivial and needs a deeper examination. On the other hand, there is a need of a discussion on modelling an expressing failures in stochastic process algebras in order to improve their usability and expressiveness [2] for modelling fault-tolerant distributed systems.

Unfortunately, according to our best knowledge, there has not been dedicated discussion or research directly concerning modelling failures in SPAs. This problem is still open for examination. There was however a discussion on application of TIPP [3] algebra to failure modelling in [4]. Authors do not directly show on how to model different failure models, but instead are interested in how dependability and performability can be examined, with the use of TIPP. Nevertheless, they provide some examples how some unspecified failures can be added to the system's model, by using the TIPP choice operator.

Given that, we would like to fill this gap by examining possible ways of modelling failures of fault-tolerant distributed systems in stochastic process algebras. This kind of discussion is needed for modellers who can use our directions to better express their models.

Along with the discussion, we present formal definitions that work as a base for our further presentation. We also show, by examples, how particular failures can be modelled in PEPA (*Performance Evaluation Process Algebra*), one of the most known stochastic process algebra. We additionally briefly discuss how one can model failure detectors.

One should note, the chapter does not deal with the state-space explosion problem caused by adding new states (e.g. failures). We treat this as an orthogonal problem and deal only with increasing of the expressiveness of process algebras.

The chapter is organised as follows. Section 12.2 introduces basic failure types known in distributed systems theory. In Section 12.3 the main contribution of the chapter is presented, whereas Section 12.4 provides a summary of the chapter.

12.2 Basic Definitions

12.2.1 Stochastic Process Algebras

Stochastic process algebras are formal methods used for modelling concurrent systems. Unlike legacy process algebras SPAs express not only pure behaviour of the system, but are also able to associate time with actions performed by processes. In stochastic process algebras, where time is exponentially distributed, models get transformed to a corresponding ergodic CTMC. For a CMTC to be ergodic, every state has to be positive recurrent and aperiodic [5]. Informally it means that models expressed in SPAs need to be sequential and there cannot exist any absorbing state, i.e. models cannot stop.

There are quite few prominent SPAs: MTIPP [6], EMPA [7] and PEPA [8, 9]. In this chapter we use PEPA as a language for the examples but the contribution of the article is applicable to every stochastic process algebra with CTMC as an underlaying mathematical foundation.

12.2.2 Basic Failures in Distributed Systems

From [10], a *failure* is an event that forces a failed process to make transition from a correct state to incorrect state, i.e. such that is not a part of the process's implementation. *Error* is a part of process's state that is responsible for a failure. *Fault* on the other hand is the cause of the *error*.

In a distributed system process failures can be classified by their model [11]. This article will focus on three basic models: *crash-recovery*, *crash-stop* and *omission*.

Crash-recovery failure model is the most interesting one from the performance point of view. In this model, failed process can be repaired and recovered to its correct state before the failure. The state it recovers to depends on the rollback recovery mechanism, but most commonly it will be the recent recorded state [12]. This model is interesting as it is the most frequently discussed, and used in fault-tolerant distributed systems. The motivation for this statement is straightforward — systems should be constructed in such a way to be able to restore its failed components.

Omissions are failures that result from the behaviour of a process which omits sending or receiving messages it was supposed to (according to its algorithm). This failure model is caused by communication faults, such as buffer overflows and network congestions. From the point of view of an external observer, this kind of failure results in apparent lack or inability in communication with the process. Omission failures can also be temporal and a failed process can finally recover and continue to function correctly.

Finally, a process can fail in a *crash-stop* model. The process exhibiting *crash-stop* failure will not perform any action after the moment of its failure. It will never be restored or repaired. In real life these failures can be seen in mobile ad-hoc networks, where mobile agents disconnect from a mobile distributed computation.

Processes that fail permanently can be restarted or rejuvenated. However, the newly started process will not be the same as the failed one. It will get different process ID, memory allocation, etc.

12.3 Modelling Failures

In order to create a base for the discussion and reasoning about the SPA models with failures, some formal definitions will be presented. We assume the definitions will be interpreted over labelled transitions systems, that provide a clear method of expressing states and transitions. Examples on the other hand will be presented in PEPA.

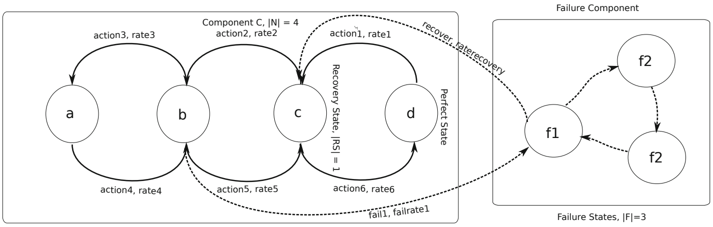


Fig. 12.1 Graphical representation of the failing component

Let us define some basic elements of the discussed model, for the sake of understanding Figure 12.1 presents a graphical view of a discussed model. The component that is failure prone will be denoted as $\mathscr{C}$, its corresponding state-space will consist of $|\mathscr{N}|$ states, belonging to $\mathscr{N}$ set. Let us also denote $S_0 \in \mathscr{N}$ as an initial state, i.e. a starting one.

The set of states that represent failure will be called *Failure States* and denoted as $\mathscr{F}$. We also assume $\mathscr{F} \notin \mathscr{N}$, what implies that we treat *Failure States* as separate from a component, for the sake of definitions.

Additionally we denote $\mathscr{T}\mathscr{F}$ as a set of all transitions from $\mathscr{N}$ to $\mathscr{F}$. We call it *failure transitions*. Similarity we define $\mathscr{T}\mathscr{R}$ as a set of all *recovery transitions* that lead from $\mathscr{F}$ to $\mathscr{C}$. This allows us to present a definition for a *general failing component*.

Definition 1 (General failing component). A component $\mathscr{C}$ is general failing component if from at least one of $\mathscr{N}$ states there is at least one transition to at least one state from $\mathscr{F}$.

Definition 1 informally says that for a component to fail, there needs to be at least one transition to states that denote failure. The definition obviously leads to a special case, where $|F| = 0$, we will call such a component a *perfect component*.

Definition 2 (Perfect component). A perfect component is such that $\mathscr{F} = \varnothing$.

Since the set F is empty, it also implies that $\mathscr{T}\mathscr{R} = \varnothing$ and $\mathscr{T}\mathscr{F} = \varnothing$.

Basing on the definitions we can construct a model of failing component using PEPA's syntax. Failures are generally probabilistic in their nature, i.e. their occurrence is dictated by a probability. In PEPA (and other stochastic process algebras) there is an operator that can express such situations — the probabilistic choice operator (denoted by +). Therefore, a general stochastic algebra model of a failing process can be expressed as: $C + F$. Here C corresponds to the failing component and F to the failure component (states that represent the behaviour after the failure). All examples presented in this chapter will fall into this model.

12.3.1 Crash-Recovery Failures

In *crash-recovery* the process is repaired and revived after the failure, see Figure 12.2 (a). It is rather straightforward that system architects are mostly interested in such a failure model, as it provides the best scenario in real life systems. After the recovery, the process continues to work from a state it was "revived" in. It is important to notice that recovery does not mean *restart*. The process is not destroyed and started as new one. It starts in the state recorded prior to the failure. The analogy would be reviving the person after his death, with all his memory from the past life restored.

Traditional failure models deal with processes as a smallest unit of abstraction. They are not interested in what state of the process the failure has occurred. However, process algebras provide a mechanism to model processes in much more grained way. Components represented by states and activities describe processes in a behavioural way. In the case of *crash-recovery* failure modelling, it needs to be decided in which states the process fails, to which states recovery actions lead, and what are the timing rates associated with all of the transitions. The same situation can be seen with the recovery action. A system does not always have to recover to the same state. In fact, a model could express a scenario where recovery is determined by the state the process failed in. Another scenario could be the need for expressing states that never fail. The situation has its real life application. Let us imagine a running system with *crash-recovery* failure model which needs to be assessed before performance tuning. The modeller could measure real rates in the system, determine to which states given processes (or servers) recover and express the probabilities of recovering in a given state. He then could parametrise a model of a system, solve it, and make decisions about what has to be changed in order to increase overall system's performance, while facing *crash-recovery* failure.

In order to define *crash-recovery* failure, we will define *recovery state* first.

Definition 3 (Recovery state). Recovery state is a state to which there is at least one transition from any of the $\mathscr{F}$ states. A set of all recovery states will be denoted as $\mathscr{R}\mathscr{S}$, where $\mathscr{R}\mathscr{S} \in \mathscr{N}$.

Basing on definition 3 we can now easily define *Crash-recovery failing component*.

Definition 4 (Crash-Recovery Failing Component). A component $\mathscr{C}$ is Crash-Recovery Failing Component if from at least one $\mathscr{F}$ state, there is at least one transition to a state in $\mathscr{RS}$ set.

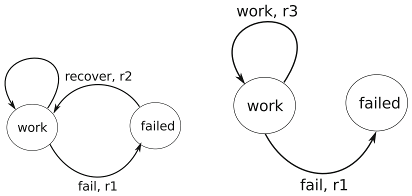
An example of a *crash-recovery failing component* is as follows:

Example 1 (Crash-recovery failure example model).

$$
\begin{array}{lll}
Process_0 \overset{def}{=} (action_0, actionRate_0).Process_1 & + & (fail_0, failRate_0).Fail_0 \\
 & \cdots & \\
Process_n \overset{def}{=} (action_n, actionRate_n).Process_0 & + & (fail_n, failRate_n).Fail_n \\
Fail_0 \overset{def}{=} (recovery_0, recoveryRate_0).Process_0 & & \\
 & \cdots & \\
Fail_n \overset{def}{=} (recovery_n, recoveryRate_n).Process_n & &
\end{array}
$$

12.3.2 Crash-Stop Failure Models

Crash-stop failure model describes process failures that cause the process to crash permanently. From the functional point of view this is a very strong assumption. Practically this is good way to model, and check properties of replicated systems, i.e. many components of one type, preferably stateless, some of them fail and never recover. In real life, such crashes can be also seen as clients that terminate or cancel their interaction with the system and never come back. From a performance modelling point of view this model has one serious implication. Components that fail at some point and never recover can be seen as terminating from the external observer point of view. In state-space world it means that a termination, or absorbing state has to be achieved, a state from which there is no way out, see Figure 12.2 (b).



(a) (b)

Fig. 12.2 Labelled transitions systems representing (a) crash-stop failure and (b) crash-recovery failure

The nature of SPA is the lack of such a termination operator, sometimes referred to as *deadlock* and denoted by *STOP* or *0* [13, 14]. The reason for that is directly emerging from the SPA's mathematical foundation, i.e. CTMC with the ergodicity property. It implies for a model that all states have to be accessible from all other states, and all states have to be visited more than once. The same requirement holds on the component level. Otherwise, the steady-state solution of the underlaying Markov process would not be possible [5], and performance measures would not be derived. Due to the ergodicity property it is obvious that a presence of a termination in any of the components would eventually deadlock the whole model. The presence of a deadlock does not however mean that no results can be achieved on the model. In such cases there is always a transient analysis of CTMCs [15]. It can be helpful to calculate state absorption probabilities. This method is of a great value for reliability analysis, yet it is of not as great value, as it comes to overall system's performance. Given the problems stated above we can now propose two ways of modelling *crash-stop* failures in SPAs. The first method deals with terminating models. Despite the fact that PEPA does not allow *STOP* operator, it is still possible to carry out transient analysis. Let us call this kind of *crash-stop* failure a *terminating crash-stop failure*. The example and a formal definition are presented below:

Definition 5 (Terminating crash-stop failing component). A component $\mathscr{C}$ is a terminating crash-stop failing component if from any state in $\mathscr{N}$ there exists at least one transition to a state from $\mathscr{F}$ and at least one of the states in $\mathscr{F}$ is a terminating one, and $|\mathscr{T}\mathscr{R}| = 0$.

The example of a *terminating crash-stop failing component* is as follows:

Example 2. Terminating crash-stop failing component

$$\begin{aligned}
Process_0 &\overset{def}{=} (action_0, actionRate_0).Process_1 + (fail_0, failRate_0).STOP \\
&\qquad\cdots \\
Process_n &\overset{def}{=} (action_n, actionRate_n).Process_0 + (fail_n, failRate_n).STOP
\end{aligned}$$

The second method of modelling *crash-recovery* failures is more compatible with the semantics of fail-stop failure, as known from distributed systems theory. As we can recall, *crash-stop* failure causes a process to vanish. This, however does not mean it cannot be restarted. The understanding of the difference between restart and recovery is crucial for our example. Upon recovery the process is revived to its *state* prior the crash, i.e. its PID, name, owner and etc. remain the same. In the case of restart the process is also recovered, but all the properties that identify its uniqueness are reset. In this case we talk about restart. If we look at *crash-stop* in that way, we can assume that *crash-stop* failure can be modelled as a special case of *crash-recovery* failure, where process after recovery is always restarted to its initial state. Let us call this kind of *crash-stop* failure a *restarting crash-stop failure*. If we denote the initial state as S_0, we can define this kind of component, as:

Definition 6 (Restarting crash-stop failing component). A component $\mathscr{C}$ is restarting crash-stop failing component if from any state in set $\mathscr{F}$ there is at least one transition tr to the S_0 state.

The example of a *restarting crash-stop failing component* is as follows:

Example 3. Restarting crash-stop failure component

$$\begin{array}{lll} Process_0 \overset{def}{=} (action_0, actionRate_0).Process_1 & + & (fail_0, failRate_0).Fail_0 \\ & \ldots & \\ Process_n \overset{def}{=} (action_n, actionRate_n).Process_0 & + & (fail_n, failRate_n).Fail_n \\ Fail_0 \quad \overset{def}{=} (restart_0, restartRate_0).Process_0 & & \\ & \ldots & \\ Fail_n \quad \overset{def}{=} (restart_n, restartRate_n).Process_0 & & \end{array}$$

12.3.3 Omission Model

In the case of *omission* failure model, not only pure states and transitions have to be taken under consideration but also what they represent, i.e. their semantics. A process fails in *omission* model if it omits steps of its algorithm responsible for sending or receiving messages. In this case such a failure can be modelled by adding transitions from states representing communication subsystem of a component to states representing a failure. In PEPA there is no direct way of modelling single messages, but what we are in fact interested in, is the impact *omission* failure may have on the system. We believe that from this point of view this kind of failure is equal to *crash-recovery* failure model. Let us take under consideration two situations that back this thesis up. The first situation is when a process P occasionally omits sending or receiving a message with some probability p. The time during the process is failed is equal to t, denoting the period of time the process omits messages. Between these, recurrent time periods, the process can be seen as correct. Another situation is when a process has a flaw that forces the process, with some probability q, to omit all messages during some arbitrary time period T, starting from time point t_0. Afterwards the process becomes correct again. From a performance point of view, these two situations are equal, as they have the same, recurring impact on the system's performance. Given the recurrent nature of PEPA components, it means the process will return to either states where it omits single messages or, in the second case, to the states where it omits all messages during some longer period of time.

Below are two examples how *omission* failure can be modelled by adding transition to a failure state, either in a component that represents a process or in a component that represents a channel.

Example 4 (Omission failure (process)).

$$\begin{array}{l} Client \overset{def}{=} (prepare, p).Client_1 \\ Client_1 \overset{def}{=} (send, s).Client + (omit, f1).Client \end{array}$$

Example 5 (Omission failure (channel)).

$$\begin{array}{l} Channel \overset{def}{=} (send, \top).Channel_1 + (recv, \top).Channel_1 + (omit, f1).Channel \\ Channel_1 \overset{def}{=} (propagate, prop1).Channel + (lost, f1).Channel_2 \\ Channel_2 \overset{def}{=} (recovery, rec1).Channel \end{array}$$

The example shows two components, one of them is a process called *Client*, the second one is some hypothetical channel. Client first prepares a message and proceeds to the next state, where it can either send it or fail to do so (omit). After that it can return to the initial state where another message can be prepared.

The Channel component can either send a message or receive one. It can also omit a message. After the omission the Channel resets itself to perform another send and receive action. In the case when the message was sent, the Channel tries to propagate it. During this step the message can also be lost. This kind of omission failure puts Channel into the state where some recovery action can take place and the Channel can be reset so it can resend the message.

12.3.4 Modelling Failure Detectors

A modeller who is interested in expressing failures would probably be as much interested in expressing mechanisms that deal with detecting them, i.e. failure detectors (FD) [11, 16]. Failure detector is an oracle that can guess whether the process is correct or incorrect [17]. Failure detector can be wrong, i.e. it can guess the state of the process wrong, we then say that a failure detector is not perfect. Perfect failure detector, on the other hand, is such that always guesses the state of the process right.

There are basically two types of failure detectors, as far as their information acquisition model is concerned. Push failure detectors are passive, they rely on the monitored process to periodically send the status information by itself. Pull FD on the other hand periodically probes the process to acquire the answer whether the process is dead or alive. Thus it can be said to be active, as it actively polls (or interrogates) process to check its state.

Push FD can be modelled by creating a failure detector component that cooperates passively with the failing component. Let us go back to the example of the restarting crash-stop failure component. The example can be easily extended to express failure detector by adding FD component. Since in the failing component there are two actions denoting failing ($fail_0$ and $fail_1$), failure detector needs to cooperate on these actions with the component. Whenever they are enabled, FD component passively synchronises on them and proceeds to state $FD1$, where detection of a failure can be expressed. The failure detector does not have impact on the performance of the failing component, because it cooperates passively (the $\top$ rate), i.e. the rate of the synchronised $fail_*$ action is determined completely by the failing component. It is worth mentioning that example below presents a perfect failure detector, i.e. such that always detects the failure correctly. It should be also possible to model an imperfect failure detector by using choice operator to denote probability of failing to detect the failure.

Example 6. Perfect push failure detector model

$$\begin{array}{lll}
State_0 \overset{def}{=} & (action_0, actionRate_0).State_1 & + \ (fail_0, failRate_0).Fail_0 \\
State_1 \overset{def}{=} & (action_1, actionRate_1).State_0 & + \ (fail_1, failRate_1).Fail_1 \\
Fail_0 \overset{def}{=} & (restart_0, restartRate_0).State_0 & \\
Fail_1 \overset{def}{=} & (restart_1, restartRate_1).State_0 & \\
FD0 \overset{def}{=} & (fail_0, \top).FD1 + (fail_1, \top).FD1 & \\
FD1 \overset{def}{=} & (failureDetected, detectRate).FD0 & \\
State_0 \underset{fail_0, fail_1}{\bowtie} & FD0 &
\end{array}$$

Pull FD periodically checks the status of the process by polling it. In order to express such a behaviour, one can still use cooperation on the action denoting failing. The trick is to express periodical behaviour of the polling mechanism. Therefore the cooperation needs to be preceded by a state that denotes time between probing. The example below presents the component of a pull failure detector. As in the case of push FD, pull FD detects failure when it synchronises on any of the fail actions. Otherwise it periodically repeats action probe followed by wait that together denote the time needed to send and receive the probe request, as well as time that passes between them.

Example 7. Perfect pull failure detector model

$$\begin{array}{ll}
FD0 \overset{def}{=} & (probe, probeRate).FD2 + (fail_0, \top).FD1 + (fail_1, \top).FD1 \\
FD1 \overset{def}{=} & (failureDetected, detectRate).FD0 \\
FD2 \overset{def}{=} & (wait, waitRate).FD0 \\
State_0 \underset{fail_0, fail_1}{\bowtie} & FD0
\end{array}$$

12.4 Conclusions

The chapter presented a problem of failures modelling in stochastic process algebras. We described a valid discussion and examples how *crash-recovery*, *crash-stop* and *omission* failure can be expressed. As an algebra of choice we used PEPA. Moreover, we proposed formal definitions of stochastic process algebra components that fail in the above failure models. The definitions have been expressed in terms of states and transitions. We showed that some of the failures that seemed to be not possible to be modelled (*crash-stop*) due to ergodicity of CTMC, yet can be expressed. On the other hand we showed an example how *omission* failure can be modelled as *crash-recovery* failure, when one takes its semantics under consideration.

Concluding the work, we think that further examination of modelling other primitives of distributed systems should be examined. This concerns channel types, nodes and communication methods. Additionally there is still a need of expressing *arbitrary* failure that was not part of this chapter. We would also like to examine ways of automatic failure injection into models. These problems will be the direction of our further work.

References

1. Clark, A., Gilmore, S., Hillston, J., Tribastone, M.: Stochastic Process Algebras. In: Bernardo, M., Hillston, J. (eds.) SFM 2007. LNCS, vol. 4486, pp. 132–179. Springer, Heidelberg (2007)
2. Bernardo, M., Donatiello, L., Ciancarini, P.: Stochastic process algebra: From an algebraic formalism to an architectural description language. In: Performance Evaluation of Complex Systems: Techniques and Tools, pp. 173–182 (2002)
3. Goetz, N., Herzog, U., Rettelbach, M.: Multiprocessor and distributed system design: The integration of functional specification and performance analysis using stochastic process algebras. In: Performance Evaluation of Computer and Communication Systems, pp. 121–146 (1993)
4. Herzog, U., Mertsiotakis, V.: Stochastic process algebras applied to failure modelling. Herzog and Rettelbach, 107–126 (1994)
5. Feller, W.: An introduction to probability and its applications, New York (1958)
6. Hermanns, H., Herzog, U., Katoen, J.-P.: Process algebra for performance evaluation. Theoretical Computer Science 274(1-2), 43–87 (2002)
7. Bernardo, M., Gorrieri, R.: A tutorial on empa: A theory of concurrent processes with nondeterminism, priorities, probabilities and time. Technical report, University of Bologna (1996)
8. Hillston, J.: Tuning systems: From composition to performance. The Computer Journal 48(4), 385 (2005)
9. Hillston, J.: A compositional approach to performance modelling. Cambridge Univ. Pr. (1996)
10. Avizienis, A., Laprie, J.C., Randell, B., Landwehr, C.: Basic concepts and taxonomy of dependable and secure computing. IEEE Transactions on Dependable and Secure Computing 1(1), 11–33 (2004)
11. Guerraoui, R., Rodrigues, L.: Introduction to Reliable Distributed Programming. Springer (2006)
12. Elnozahy, E.N., Alvisi, L., Wang, Y.-M., Johnson, D.B.: A survey of rollback-recovery protocols in message-passing systems. ACM Comput. Surv. 34, 375–408 (2002)
13. Thomas, N., Bradley, J.: Terminating processes in PEPA. In: University of Leeds, Citeseer (2001)
14. Hennessy, M.: Algebraic theory of processes. The MIT Press, Cambridge (1988)
15. Baier, C., Haverkort, B., Hermanns, H., Katoen, J.P.: Model Checking Continuous-time Markov Chains by Transient Analysis. In: Emerson, E.A., Sistla, A.P. (eds.) CAV 2000. LNCS, vol. 1855, pp. 358–372. Springer, Heidelberg (2000)
16. Freiling, F.C., Guerraoui, R., Kuznetsov, P.: The failure detector abstraction. ACM Comput. Surv. 43, 9:1–9:40 (2011)
17. Chandra, T.D., Toueg, S.: Unreliable failure detectors for reliable distributed systems. J. ACM 43, 225–267 (1996)

Chapter 13
Experimental Evaluation of Resampling Combined with Clustering and Random Oracle Using Genetic Fuzzy Systems

Tadeusz Lasota, Zbigniew Telec, Bogdan Trawiński, and Grzegorz Trawiński

Abstract. The ensemble methods combining resampling techniques: cross-validation, repeated holdout, and bootstrap sampling with clustering and random oracle using a genetic fuzzy rule-based system as a base learning algorithm were developed in Matlab environment. The methods were applied to the real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transactions. The computationally intensive experiments were conducted aimed to compare the accuracy of ensembles generated by the proposed methods with different number of clusters or random oracle subsets. The statistical analysis of results was made employing nonparametric Friedman and Wilcoxon statistical tests.

13.1 Introduction

We have been performing extensive investigation to select appropriate machine learning methods which would be useful for developing an automated system to assist with real estate appraisal designed for information centres maintaining

Tadeusz Lasota
Wrocław University of Environmental and Life Sciences,
Department of Spatial Management, Wrocław, Poland
e-mail: tadeusz.lasota@up.wroc.pl

Zbigniew Telec · Bogdan Trawiński Wrocław
Wrocław University of Technology, Institute of Informatics, Wrocław, Poland
e-mail: {zbigniew.telec,bogdan.trawinski}@pwr.wroc.pl

Grzegorz Trawiński
Wrocław University of Technology, Faculty of Electronics, Wrocław, Poland
e-mail: grzegorztrawinski@wp.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 131–142.
springerlink.com

cadastral systems in Poland. So far, we have examined several methods to construct regression models to assist with real estate appraisal: evolutionary fuzzy systems, neural networks, decision trees, and statistical algorithms using MATLAB, KEEL, RapidMiner, and WEKA data mining systems [6], [11], [14]. A good performance revealed evolving fuzzy models applied to cadastral data [15], [18]. We studied also ensemble models created applying various weak learners and resampling techniques [10], [16], [17]. In this chapter we enhance the resampling methods [1], [2], [19] by combining them with dynamic regressor selection. This approach for classification problems was employed in [13], [22]. At the training stage we partition a given training dataset into few disjoint groups using clustering or random oracle approaches and build models over individual groups. During the predicting phase only one model is selected from among all available ones to produce the output. Given a testing instance as the selector the nearest cluster center or random oracle is used. Having all predictions given by the selected models we compute the final result of a given resampling technique as the arithmetic mean of model accuracies.

For clustering one of the well-known algorithms was used, namely K-Means algorithm, which belongs to centroid based methods. K-Means was described in numerous textbooks and scientific works [7],[8],[9]. It partitions a set of objects, in our case training instances, into k disjoint clusters with low intra-cluster distances and high inter-cluster distances. For the determination of the optimal number of clusters and evaluation of the quality of the partitions many validity indices were proposed and examined. Among them Davies–Bouldin and Dunn indices belong to the most popular methods devoted to crisp partitions [5], [20].

Random oracle is a relatively new method of ensemble design devised by Kuncheva and Rodriguez and applied to classification and regression problems [12], [21]. They used linear random oracle which divided the space into two subspaces using a hyperplane. To build the oracle two different training instances were randomly selected, and then, each remaining training instance was assigned to the subspace determined by the closer selected instance. In the training phase, two models were built, each over one of the two so obtained subsets. In the prediction phase, for one test instance only one of the two models was used. The authors argued that this approach adds an extra diversity to the ensemble and therefore allows for high accuracy of the individual ensemble members.

The idea of our automated valuation system assumes a data driven modeling framework for premises valuation developed with the sales comparison method. The main advantage of data driven models is that they can be automatically generated from given datasets and, therefore, save a lot of time and financial supply. Sometimes, it is necessary to use this kind of models due to the high complexity of the process to be modeled. It was assumed that the whole appraisal area, that means the area of a city or a district, is split into sections of comparable property attributes. The architecture of the proposed system is shown in Fig. 13.1. The appraiser accesses the system through the internet and chooses an appropriate section and input the values of the attributes of the premises being evaluated into the system, which calculates the output using a given model. The final result, as a suggested value of the property, is sent back to the appraiser.

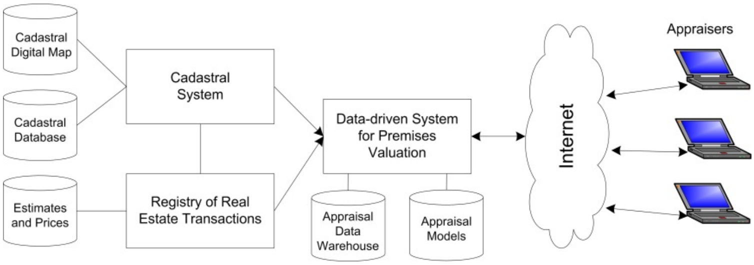


Fig. 13.1 Schema of automated data-driven system for property valuation

13.2 Methods Used in Experiments

Three following resampling methods were applied in the experiments and combined with clustering techniques: 10-fold cross-validation (CV10x1), holdout with the split into training and test sets in the proportion 80% to 20% and repeated 10 times (HO80x10), and finally bootstrap sampling of 100% instances from base datasets with replacement to form a training set and using base datasets as test sets, repeated also 10 times (BS100x10). Thus, in each case an ensemble comprising 10 component models was created and as an output the arithmetic mean of model accuracies was computed. They were combined with clustering or random oracle, which at the training stage allowed for the partition training sets into few disjoint groups, and then develop genetic fuzzy systems over individual clusters or subsets. In turn, at the predicting stage for subsequent test instances only one model was selected from among all available ones to provide the predicted output value. The procedure is described in Algorithm 1 and schemata of experiments in the case of clustering are illustrated in Fig. 13.2-13.4.

Algoritm 1. Pseudocode of resampling methods combined with clustering or random oracle

Given:

- R: number of repetitions, i.e. splits of a base dataset in a resampling techniques
- x_i, x_j: instances from a training or test sets (vector of input values)
- K: number of clusters or random oracle subsets for a given training set
- C_{ik}, C_{ik}^{RO}: a cluster, a random oracle subset respectively
- c_{ik}: a cluster centre
- x_{ik}^{RO} : a random oracle instance
- $|C_{ik}|$, $|C_{ik}^{RO}|$: cardinality, i.e. the number of elements in a cluster or subset
- N_{min}: minimal number of elements in a cluster or subset
- DI_l: Dunn's Index of the l-th partition into clusters
- $d(x_j, c_{ik})$ – Euclidean distance between x_j and c_{ik}
- GFS: genetic fuzzy system used as a base learning algorithm
- FIS(x): value predicted by a fuzzy model GFS for an instance x
- MSE: mean squared error

Resampling

Split randomly a base dataset into R pairs of training T_i and test sets S_i according to the resampling schema (i=1,2,…,R)

For clustering

Training Phase

```
For i=1,2,   ,R
  • Take the training set $T_i$
  • For l=1,2,…,L
      • Apply the K-Means algorithm to partition $T_i$ into K clusters $C_{ik}$, represented by cluster centres $c_{ik}$ (k=1,..K)
      • If any $|C_{ik}|<N_{min}$ discard the partition
      • Compute $DI_l$ for l-th partition
  • EndFor
  • Select the partition with the highest $DI_l$
  • Build $GFS_{ik}$ over each $C_{ik}$ of the selected partition
EndFor
```

Predicting Phase

```
For i=1,2,   ,R
    • For each instance $x_j$ from a test set, select one fuzzy model $GFS_{ik}$ for which $d(x_j, c_{ik})$ is minimal
    • Let $FIS_{ij}(x_j)$ be the value predicted by the selected fuzzy model $GFS_{ik}$
    • Compute the $MSE_i$ using all predicted and actual values
EndFor
```

For Random Oracle

Training Phase

```
For i=1,2,   ,R
  • Take the training set $T_i$
  • from $T_i$ draw randomly K instances and call them random oracle instances $x_{ik}^{RO}$ (k=1,2,…,K)
  • Each instance $x_j$ from a training dataset, assign to the random oracle subset $C_{ik}^{RO}$ for which $d(x_j, x_{ik}^{RO})$ is minimal
  • If any $|C_{ik}^{RO}|<N_{min}$ discard the partition and repeat from drawing random oracle instances
  • Build $GFS_{ik}$ over each random oracle subset $C_{ik}^{RO}$
EndFor
```

Predicting Phase

```
For i=1,2,   ,R
    • For each instance $x_j$ from a test set $S_i$, select one fuzzy model $GFS_{ik}$ for which $d(x_j, x_{ik}^{RO})$ is minimal
    • Let $FIS_{ij}(x_j)$ be the value predicted by the selected fuzzy model $GFS_{ik}$
    • Compute the $MSE_i$ using all predicted and actual values
EndFor
```

Computing the output

The final output is computed as $MSE = \frac{1}{R}\sum_{i=1}^{R} MSE_i$

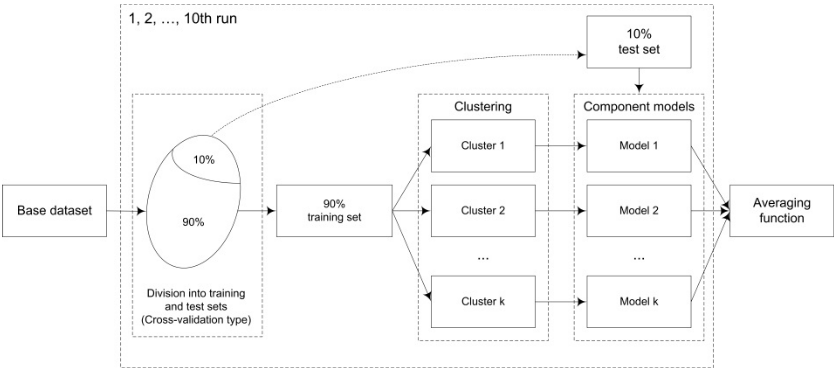


Fig. 13.2 Outline of experiment with clustering within the Cross-validation frame (CV10x1)

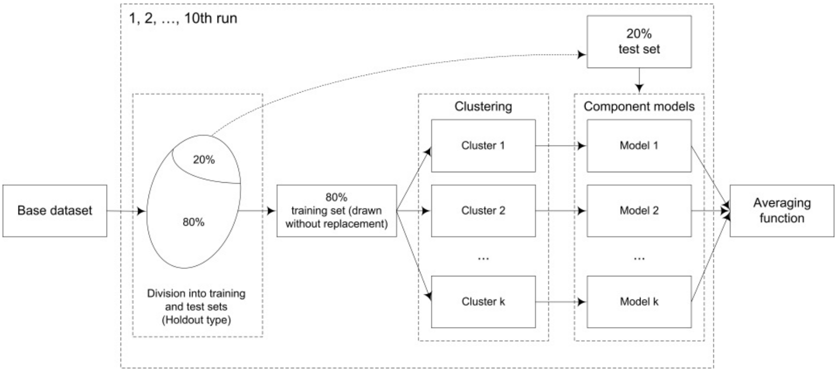


Fig. 13.3 Outline of experiment with clustering within the Holdout frame (HO80x10)

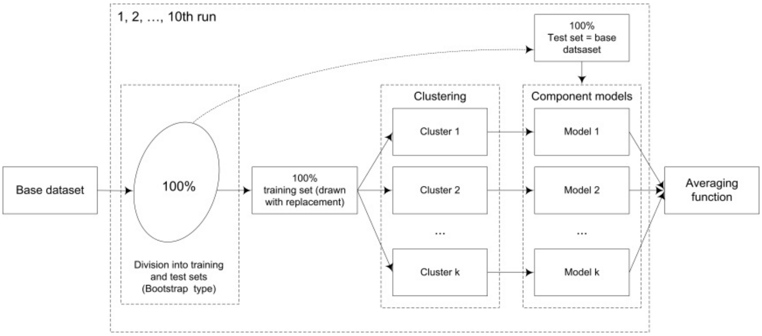


Fig. 13.4 Outline of experiment with clustering within the Bootstrap frame (BS100x10)

13.3 Experimental Setup

Real-world data used in experiments was drawn from an unrefined dataset containing above 50 000 records referring to residential premises transactions accomplished in one Polish big city with the population of 640 000 within eleven years from 1998 to 2008. The dataset was confined to sales transaction data of apartments built before 1997 and where the land was leased on terms of perpetual usufruct. Hence, the final dataset counted 5303 records. Four following attributes were pointed out as price drivers by our experts who were professional appraisers: usable area of a flat (*Area*), age of a building construction (*Age*), number of storeys in the building (Storeys), and the distance of the building from the city centre (*Centre*), in turn, price of premises (*Price*) was the output variable.

Due to the fact that the prices of premises change substantially in the course of time, the whole 11-year dataset cannot be used to create data-driven models. In order to obtain comparable prices it was split into subsets covering individual years. Then the prices of premises were updated according to the trends of the value changes over 11 years. Starting from the beginning of 1998 the prices were updated for the last day of subsequent years. The trends were modelled by polynomials of degree three. We might assume that one-year datasets differed from each-other and might constitute different observation points to compare the accuracy of ensemble models in our study. The sizes of one-year datasets are given in Table 13.1. As a performance function the mean square error (MSE) was used, and as aggregation functions arithmetic averages were employed. Each input and output attribute was normalized using the min-max approach.

Table 13.1 Number of instances in one-year datasets

1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
269	477	329	463	530	653	546	580	677	575	204

Table 13.2 Parameters of GFS used in experiments

Fuzzy system	Genetic Algorithm
Type of fuzzy system: Mamdani	Chromosome: rule base and mf, real-coded
No. of input variables: 5	Population size: 50
Type of membership functions (mf): triangular	Fitness function: MSE
No. of input mf: 3	Selection function: tournament
No. of output mf: 5	Tournament size: 4
No. of rules: 15	Elite count: 2
AND operator: prod	Crossover fraction: 0.8
Implication operator: prod	Crossover function: two point
Aggregation operator: probor	Mutation function: custom
Defuzzyfication method: centroid	No. of generations: 100

In turn, in GFS approach for each input and output variable three triangular and trapezoidal membership functions were automatically determined by the symmetric division of the individual attribute domains. The evolutionary optimization process combined both learning the rule base and tuning the membership functions using real-coded chromosomes. Similar designs are described in [3], [4], [11]. The parameters of the architecture of fuzzy systems as well as genetic algorithms are listed in Table 13.2.

13.4 Experimental Results

The performance of CV10x1, HO80x10, BS100x10 models created using genetic fuzzy systems (GFS) over 1-5 clusters and 1-5 random oracle subsets is shown in Figures 13.5-13.7 and 13.8-13.10, respectively. The statistical analysis of the results was carried out with non-parametric Friedman and Wilcoxon tests performed in respect of MSE values of all ensembles built over 11 one-year datasets. Average ranks of individual ensembles provided by Friedman tests are shown in Table 13.3 and 13,5 for clustering and random oracle, respectively, where the lower rank value the better model. The results are statistically signiificant for p-value<0.05.

In Table 13.4 and 13.6 the results of nonparametric Wilcoxon signed-rank test to pairwise comparison of the model performance are presented for clustering and random oracle, respectively. The zero hypothesis stated there were not significant differences in accuracy, in terms of MSE, between given pairs of models. In both tables + denotes that the model in the row performed significantly better than, – significantly worse than, and ≈ statistically equivalent to the one in the corresponding column, respectively. In turn, / (slashes) separate the results for individual resampling methods. The significance level considered for the null hypothesis rejection was 5%. Only for bagging the differences in accuracy of the ensembles created using different number of groups revealed statistical significance, where the bigger number of clusters or random oracle subsets the lower values of MSE.

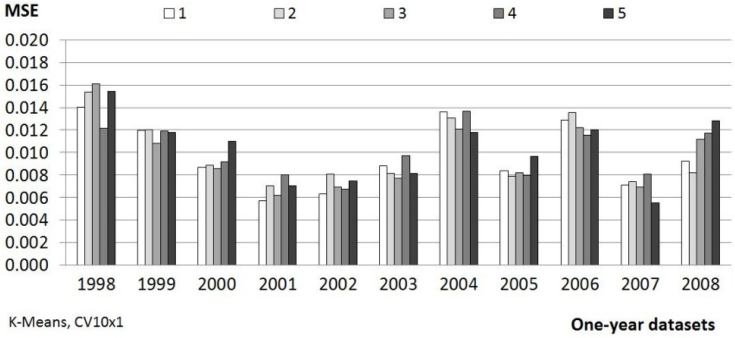


Fig. 13.5 Performance of models built using CV10x1 combined with K-Means

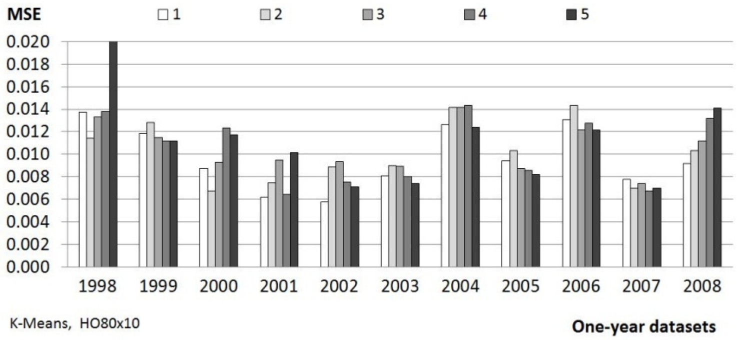


Fig. 13.6 Performance of models built using HO80x10 combined with K-Means

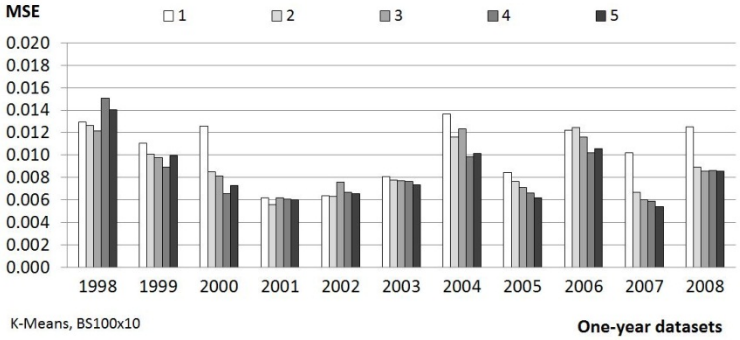


Fig. 13.7 Performance of models built using BS100x10 combined with K-Means

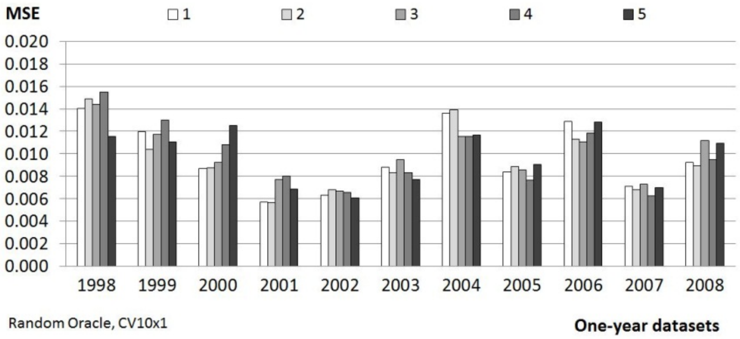


Fig. 13.8 Performance of models built using CV10x1 combined with Random Oracle

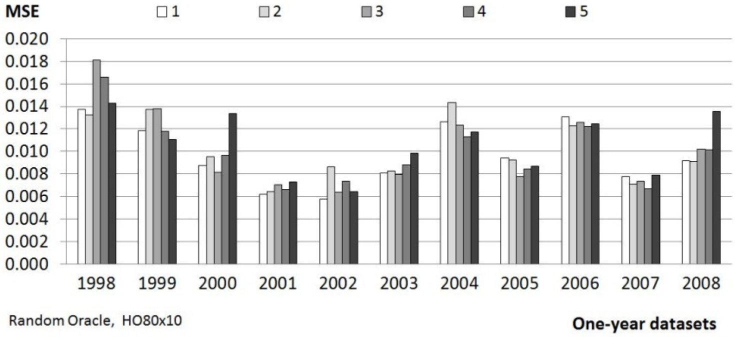


Fig. 13.9 Performance of models built using HO80x10 combined with Random Oracle

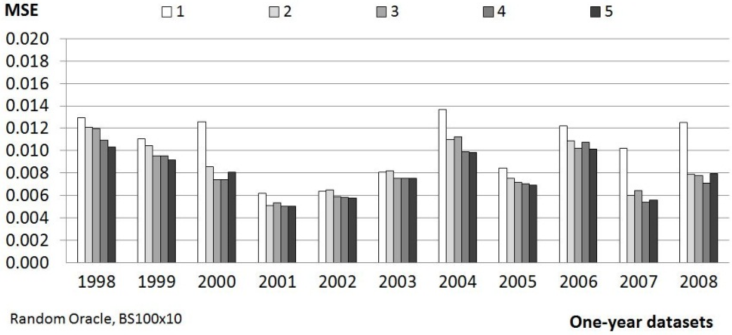


Fig. 13.10 Performance of models built using BS100x10 combined with Random Oracle

Table 13.3 Results of Friedman test for different number of clusters in K-Means

CV10x1		HO80x10		BS100x10	
# gr.	Ranking	# gr.	Ranking	# gr.	Ranking
3	2.36	1	2.73	5	2.09
1	2.82	5	2.73	4	2.27
5	3.18	4	2.91	3	2.91
2	3.27	3	3.27	2	3.27
4	3.36	2	3.36	1	4.45
p-value	0.56087	p-value	0.80879	p-value	0.00355

Table 13.4 Results of Wilcoxon tests for different number of clusters in K-Means (CV10x1/HO80x10/BS100x10)

# groups	1	2	3	4	5
1		~ / ~ / -	~ / ~ / –	~ / ~ / –	~ / ~ / –
2	~ / ~ / +		~ / ~ / ~	~ / ~ / ~	~ / ~ / ~
3	~ / ~ / +	~ / ~ / ~		~ / ~ / ~	~ / ~ / ~
4	~ / ~ / +	~ / ~ / ~	~ / ~ / ~		~ / ~ / ~
5	~ / ~ / +	~ / ~ / ~	~ / ~ / ~	~ / ~ / ~	

Table 13.5 Results of Friedman test for different number of subsets in Random Oracle

CV10x1		HO80x10		BS100x10	
# gr.	Ranking	# gr.	Ranking	# gr.	Ranking
2	2.73	4	2.64	5	1.73
1	2.91	1	2.82	4	1.91
5	2.91	2	2.91	3	2.73
4	3.09	3	3.00	2	3.82
3	3.36	5	3.64	1	4.82
p-value	0.90703	p-value	0.63652	p-value	0.00000

Table 13.6 Results of Wilcoxon tests for different number of subsets in Random Oracle (CV10x1/HO80x10/BS100x10)

# groups	1	2	3	4	5
1		~ / ~ / –	~ / ~ / –	~ / ~ / –	~ / ~ / –
2	~ / ~ / +		~ / ~ / ~	~ / ~ / –	~ / ~ / –
3	~ / ~ / +	~ / ~ / ~		~ / ~ / ~	~ / ~ / ~
4	~ / ~ / +	~ / ~ / +	~ / ~ / ~		~ / ~ / ~
5	~ / ~ / +	~ / ~ / +	~ / ~ / ~	~ / ~ / ~	

13.5 Conclusions

The ensemble methods combining resampling techniques: cross-validation, repeated holdout, and bootstrap sampling (i.e. bagging) with dynamic regressor selection using a genetic fuzzy rule-based system as a base learning algorithm were developed in Matlab environment. The dynamic regressor selection consisted in employing clustering or random oracle, which in the training phase allowed for the partition training sets into few disjoint groups, and then develop genetic fuzzy systems over individual clusters or subsets. Next, in the predicting phase for subsequent test instances only one model was selected from among all available ones to provide the predicted output value. Given a test instance as the selector the nearest cluster center or random oracle instance was used. It is argued that this approach adds an extra diversity to the ensemble and therefore allows for high accuracy of the individual ensemble members. Moreover, the approach provides a learning algorithm with more uniform training instances what may result in increased accuracy of the models generated.

The computationally intensive experiments aimed to compare the performance of proposed methods over real-world data taken from a cadastral system with different numbers of clusters and random oracle subsets were conducted. The statistical analysis of results was made employing nonparametric Friedman and Wilcoxon statistical tests. The overall results of our investigation are as follows. The differences in accuracy of the ensembles created using different number of clusters or random oracle subsets turned out to be statistically significant only in the case of bootstrap resampling (i.e, bagging), where the bigger number of groups the lower values of mean squared error. Further investigations into resampling methods combined with other partitioning methods such as stratification and using other base learning algorithms such as neural networks or decision trees are planned. Benchmark regression datasets preprocessed with instance and feature selection algorithms will be used and the resistance of the methods to noised data will be also examined.

Acknowledgments. This work was partially supported by the Polish National Science Centre under grant no. N N516 483840.

References

1. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 123–140 (1996)
2. Bühlmann, P., Yu, B.: Analyzing bagging. Annals of Statistics 30, 927–961 (2002)
3. Cordón, O., Gomide, F., Herrera, F., Hoffmann, F., Magdalena, L.: Ten years of genetic fuzzy systems: current framework and new trends. Fuzzy Sets and Systems 141, 5–31 (2004)
4. Cordón, O., Herrera, F.: A Two-Stage Evolutionary Process for Designing TSK Fuzzy Rule-Based Systems. IEEE Tr. on Sys., Man, and Cyb. -Part B 29(6), 703–715 (1999)
5. Dunn, J.C.: Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4(1), 95–104 (1974)
6. Graczyk, M., Lasota, T., Trawiński, B.: Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS(LNAI), vol. 5796, pp. 800–812. Springer, Heidelberg (2009)
7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kauffman (2006)
8. Hartigan, J.A., Wong, M.A.: A K-Means Clustering Algorithm. Applied Statistics 28(1), 100–108 (1979)
9. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer (2009)

10. Kempa, O., Lasota, T., Telec, Z., Trawiński, B.: Investigation of Bagging Ensembles of Genetic Neural Networks and Fuzzy Systems for Real Estate Appraisal. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNCS(LNAI), vol. 6592, pp. 323–332. Springer, Heidelberg (2011)
11. Król, D., Lasota, T., Trawiński, B., Trawiński, K.: Investigation of Evolutionary Optimization Methods of TSK Fuzzy Model for Real Estate Appraisal. International Journal of Hybrid Intelligent Systems 5(3), 111–128 (2008)
12. Kuncheva, L.I., Rodríguez, J.J.: Classifier Ensembles with a Random Linear Oracle. IEEE Transactions on Knowledge and Data Engineering 19(4), 500–508 (2007)
13. Kuncheva, L.I.: Switching between Selection and Fusion in Combining Classifiers: An Experiment. IEEE Trans. Systems, Man, and Cybernetics, Part B 32(2), 146–156 (2002)
14. Lasota, T., Mazurkiewicz, J., Trawiński, B., Trawiński, K.: Comparison of Data Driven Models for the Validation of Residential Premises using KEEL. International Journal of Hybrid Intelligent Systems 7(1), 3–16 (2010)
15. Lasota, T., Telec, Z., Trawiński, B., Trawiński, K.: Investigation of the eTS Evolving Fuzzy Systems Applied to Real Estate Appraisal. Journal of Multiple-Valued Logic and Soft Computing 17(2-3), 229–253 (2011)
16. Lasota, T., Telec, Z., Trawiński, G., Trawiński, B.: Empirical Comparison of Resampling Methods Using Genetic Fuzzy Systems for a Regression Problem. In: Yin, H., Wang, W., Rayward-Smith, V. (eds.) IDEAL 2011. LNCS, vol. 6936, pp. 17–24. Springer, Heidelberg (2011)
17. Lasota, T., Telec, Z., Trawiński, G., Trawiński, B.: Empirical Comparison of Resampling Methods Using Genetic Neural Networks for a Regression Problem. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011, Part II. LNCS(LNAI), vol. 6679, pp. 213–220. Springer, Heidelberg (2011)
18. Lughofer, E., Trawiński, B., Trawiński, K., Kempa, O., Lasota, T.: On Employing Fuzzy Modeling Algorithms for the Valuation of Residential Premises. Information Sciences 181, 5123–5142 (2011)
19. Molinaro, A.N., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics 21(15), 3301–3307 (2005)
20. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. Pattern Recognition 37(3), 487–501 (2004)
21. Pardo, C., Rodríguez, J.J., Díez-Pastor, J.F., García-Osorio, C.: Random Oracles for Regression Ensembles. In: Okun, O., Valentini, G., Re, M., et al. (eds.) Ensembles in Machine Learning Applications. SCI, vol. 373, pp. 181–199. Springer, Heidelberg (2011)
22. Woods, K., Kegelmeyer, W.P., Bowyer, K.: Combination of Multiple Classifiers Using Local Accuracy Estimates. IEEE Trans. Pattern Analysis and Machine Intelligence 19(4), 405–410 (1997)

Chapter 14
Ant Colony Optimization Parameter Evaluation

Andrzej Siemiński

Abstract. The chapter addresses the problem of parameter evaluation for the Ant Colony Optimization (ACO) technique. The operation of the ACO is too complex to allow for an analytical approach to the problem of optimizing parameter setting. Therefore their values are usually chosen in an experimental way. The chapter presents an in depth analysis of the impact of the individual parameters on overall ACO performance and studies their interplay. The analyzed version of the ACO is used for solving the Travelling Salesmen Problem (TSP). Both static and dynamic versions of the problem are considered. In the dynamic environment 4 modes of route variability are studied. The chapter ends with a statistical analysis of data gathered in a sequence of experiments.

14.1 Introduction

The aim of the chapter is to discuss the problem of setting parameter values for Ant Colony Optimization (ACO) in both static and dynamic environments. The operation of the ACO is controlled by a number of parameters. The complexity of its operation makes it impossible to provide an analytical solution to the problem of optimizing their values. Therefore the values are usually chosen in an experimental manner. Most work on the area concentrates upon modifying the operation of the basic ACO engine in order to improve its performance of adopt it to new tasks. It is out belief that the analysis of the influence of individual parameters and their interplay have not received enough attention. Recent papers presented in the chapter 14.3 suggest that proper selection of parameters offers the possibility to improve the ACO performance. We are also convinced that the in depth insight into the role of individual parameters and their interplay could help to propose a specialized version of the basic algorithm. This is specially important for versions developed for the inherently complex dynamic environments.

Andrzej Siemiński
Wroclaw University of Technology, Institute of Informatics, Wrocław, Poland
e-mail: Andrzej.Sieminski@pwr.wroc.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 143–153.
springerlink.com

The chapter is organized as follows. The second section presents the basic version of the ACO to the Travelling Salesman Problem for a static environment. Its parameters and their impact on the **its** operation are studied. The section ends with the discussion of to recent papers devoted to the problem of selection proper parameter values. The third section presents the Dynamic TSP. It starts with the presentation of modifications to the basic algorithm introduced in order to handle modification of route lengths. The section continues with the presentation of 4 different types of changeable graphs that are considered in the chapter. They are more complex then the graphs presented usually in the papers and it is our belief that they cover a wide application area. It concludes with the discussion f parameter interplay in the dynamic environment. The next section describes the data gathering process. Statistical analysis of obtained results and their comparison to theoretical deliberations from the third section. The chapter concludes with 5th section.

14.2 Basic ACO Algorithm for Static TSP

The Ant Colony Optimization is a popular meta-heuristic for solving combinatorial optimization problems. It was described for the first time by M. Dorigo in his PhD thesis [1] in 1992 who up to now remains one of the key researcher on the field. The ACO was inspired by the behavior of real ants and its first application area was the Travelling Agent Problem (TSP). The problem consists in finding a shortest route that connects all cities on a map provided that each city is visited only once. In what follows the cites and the map are represented by nodes of a weighted, symmetric graph. The TSP has been proved to be a NP hard problem and the ACO is one of the heuristics used successfully to solve it. A recent and comprehensive account of the state of art of ACO is presented in [2].

An ant could be regarded as an extremely simple agent. All it could do is to move from one node to another laying a pheromone trail on its way. It is also capable of detecting its current position, remembering the nodes it has visited so far and sensing the direct distances from its current position to other nodes and also the amount of pheromone laid on them. A set of ants is a part of an Ant Colony. The colony works in iterations. At the start or each iteration the ants are placed randomly on the graph. In each step of an iteration an ant selects most valuable node that was not visited so far. The iteration stops when all cities are included in an ants' route. The Ant Colony is not just a set of ants but it also harvests the collective intelligence of individual ants. It remembers the best ant in the current iteration, the best so far ant and it performs global pheromone updating.

Another part of an Ant Colony is an Ant Graph. It is defined by a set of n nodes and two functions η and τ that specify:

- $\eta(r,t)$: the distance between two nodes r and t, a value in the range [0..1];
- $\tau(r,t)$: the amount of pheromone that resides on the path from node r to t, a positive value.

14.2.1 ACO Basic Operations

The basic operations defining ACO behavior are defined by three rules:

- State transition rule that specifies in what manner an ant selects the next node to visit.
- Local updating rule that describes the way of updating pheromone values by an ant as it finds its route.
- Global updating rule which is initialized at the and of each iteration by a colony and effects the pheromone values of the whole graph.

The formulas defining the above rules could be found in many sources e.g. in [3]. The exact operation of the formulas is controlled by 4 parameters:

Q0: probability of selecting exploitation over exploration;
α: aging factor used in the global updating rule
β: moderating factor for the utility function
ρ: aging factor in the local updating rule

The State Transition Rules selects an available (not yet visited node) using one of two algorithms: exploration or exploitation. The parameter Q0 specifies the probability of selecting the exploitation. The exploitation algorithm is a deterministic one and it selects a route which maximizes the value of the following path utility function:

$$qf(r,t) = \quad \tau(r,t) * \eta(r,t)^{\beta} \tag{14.1}$$

The exploration mode of work has a probabilistic nature and the above formula specifies the probability of using the path from r to t. the where r and t are the respectively the current and destination nodes whereas β is a parameter. The higher the value of Q0 the more predictable is an ants' behavior and the diversity of paths decreases. This could lead to finding a premature selection of a local optima.

The parameter β specifies the influence of pheromones levels which represent the collective knowledge of the ACO. The values of β are always >1 and η(r,t) <=1 therefore increasing the value of β gives more prominence to pheromone level. In the case of dynamic environments this could proof to be potentially dangerous as the pheromone levels were accumulated for not longer valid path distances.

The two aging factors determine the rate of pheromone evaporation. The α parameter is used for the global updating rule and is applied for all routes whereas the ρ parameter effects only routs taken by an individual ant. Their values are in the range from 0 to 1 and the higher the value is the more rapidly the pheromone evaporates. This could prove beneficial especially for the dynamic environments.

The analysis suggests that the finding the parameter values is rather straightforward. The experiments show however that their interplay makes the process far more complex.

14.2.2 Parameter Selection

The complexity of the ACO makes it impossible to find the optimal values of the various parameters in an analytic manner. There is not much emphasis selection process in the literature. Usually only the used values are given followed by a statement that they were selected in an experimental manner [3].

Recently a paper was published [4] that attempted to make the process more refined. It was inspired by a concepts taken from Evolutionally Programming (EP) and Simulated Annealing (SA). In the paper a coding of floating point parameter values enabled crossover and much emphasis was put on mutation. An adaptation of Artificial Annealing schema was also used to gradually limit the scope genetic modifications. As a result the identified parameter values produced results better then the default parameter values. One disturbing factor was that the results although acceptable in themselves did no show much evidence on grouping around certain values. We are going to return to the phenomena in the chapter 14.5.

The Table 14.1 summarizes the parameters their ranges of values tested in the study and their recommended by U. Chirico values.

Table 14.1 ACO parameter description

Name	Description	Suggested Value	Tested Range
Q0	Probability of selecting exploitation over exploration	0.8	0.10-0.99
α	Aging factor used in the global updating rule	0.1	0.01-0.5
β	Moderating factor for the cost measure function	2.0	1.0-4.0
ρ	Aging factor in the local updating rule	0.1	0.01-0.5

A combination of values of the 4 above parameters is called a parameter test set (PTS). In the study the values in the PTS were set by a random number generator that produced uniform values from the appropriate ranges.

14.3 Dynamic TSP

In recent years we witness a growing interest in addressing dynamic optimization problems. One of the most popular research areas is the dynamic travelling salesman problem (DTSP) - a modification of the classic TSP in which the routes length are subject to change. It is regarded as one most challenging problems and difficult NP problems in computer science. The study on the area have both theoretical significance and have important practical applications. Theyinclude among others route selection for letter carriers and package routing in communication networks, goods distribution sequence.

14.3.1 Related Work

For the first time the DTSP was discussed by Psaraftis in [5]. His work however was mainly focused on problem defining, algorithm designing, performance estimation and test-bed construction and not on providing solutions.

The first attempts to modify the standard operation of ACO were concentrated upon introducing global and local reset strategies [6]. A change in the distances obviously invalidates part of accumulated pheromone levels. Global reset is easily to implement but highly computationally inefficient as it starts the optimization process once more. The local reset enables the Colony to exploit at least part of data gathered so far but requires a precise data on where the change had occurred. Both of the reset strategies and are less efficient or even entirely not useful in the case of constantly changing environments.

The alternative way to ensure ant population diversity necessary to adopt to changing environment is to implement the immigrant schemes. They consist in introducing new individuals into the current population. There are three types of immigrants: traditional random, elitism-based, and hybrid immigrants. The approach could use a long-term memory as in P-ACO [7]. A more recent paper a short-term memory is used [8]. The study reviled that different immigrants schemes are advantageous under different environmental conditions.

Studying the DTSP one has to develop the graph modification schemes. The first papers considered only a single node deletion or introduction, which makes it difficult to represents any real-life application. The previously mentioned paper [8] discusses also the benchmarking the DTSP. It introduces graphs that are modified continuously by using two basic operations: node deletion and node insertion. The operations are controlled by a random number generator. Changing route length is achieved by applying both of them to the same node.

14.3.2 Graph Generators

In what follows the distances between graphs nodes are in the range form 0.0 to 1.0. Their initial values are generated by a uniform random number generator. Contrary to some previous graphs the changes are not confined to a known in advance and rather small fragment of the whole graph. The chapter introduces a more general approach. The usual interpretation of the distance matrix is geographical distance that separates the nodes. It could be however interpreted as the average time that is necessary to move from one node to another. In this case a graph that continually allows for changes in the distance values are a far more realistic approximation of the ever changing road conditions.

For that reason a graph is replaced by a graph generator. In each iteration a generator can modify its distance matrix. In the study 4 types of such generators were considered.

- Constant Distances generator

The Constant Distances (CD) generator produces the same initial distance matrix so it is used for reference purposes and reflects the static environment.

- One Change generator

The graph generator (OC) modifies the distances during only one iteration. It is used for testing the behavior of different Ant Colonies. To maximize its impact on operation of Ant Colony the path length the modification effects the shortest distances and longest paths found in the graph. The parameters are:

- oNChange: the number of modifications. Each modification changes the originally shortest and longest distances.
- oIterNo the iteration number at which the distances are modified.

The distances are modified according to the formula: $d_{new} = 1- d_{old}$;

where d_{new} and d_{old} represent the initial and resulting distances.

The formula for distances modification should guarantee that the changes effect the best so far route and at the same time do not change significantly the average segment distance.

- Memory less generator

The modifying distances by the Memory Less (ML) generator is controlled by a memory less information source. The scope of changes is random. The distances is are modified after each iteration. It has two parameters:

- pScope – the maximal range of a change, it has a value in the range from 0.0 to 1.0
- pChange – the probability of a single distance change. In the process the length may increase or decrease with equal probability.

The new distances are computed according to the below formulas.

- $d_{new} = d_{old}$ *(1-pScope*Rand()) for decreasing the distance
- $d_{new} = d_{old}$ +(1- d_{old})*pScope*Rand() for increasing the distance
- where d_{new} and d_{old} refer to the old and new distance value and Rand() is a function that produces random numbers in the range from 0.0 to 1.0.

- Two state generator

The operation of the Two State Generator (TS) is controlled by a Markov information source with two internal states named A and B and transition matrix T. In each of them it works as a ML generator. It has two parameters are:

- T– a 2x2 matrix used to specify the probabilities of moving from one state to another;
- pScope – a matrix of two values from the range [0.0..1.0]. They define the scope of changes in the respective states. The new distances are calculated according to the above formulas for the ML generator.

Table 14.2 Graph generators used in the study

Type	Code	Parameters	Comments
CD	CD	None	It is used as a reference
OC	OCa	nChange=1, iterNo=30	A single change occurs before the Colony converges
OC	OCb	nChange==10 iterNo=30	Substantial number of changes before the Colony converges
OC	OCc	nChange=1, iterNo=300	A Single change after the Colony converges
OC	OCd	nChange=10, iterNo=300	Substantial number of changes after the Colony converges
ML	MLa	pChange=0.001, pScope=0.1	Minor and not frequent distances changes
ML	MLb	pChange=0.01, pScope=0.1	Minor and frequent distances changes
ML	MLc	pChange=0.001, pScope=0.3	Major and not frequent distances changes
ML	MLd	pChange=0.01, pScope=0.3	Major and frequent distances changes
TS	TSa	Tran={{0.999, 0.001}{0.4, 0.6}}, pScope={0.0; 0.1}	Long periods of stability with shorter periods of small distance changes
TS	TSb	Tran={{0.999, 0.001}{0.2, 0.8}}, pScope={0.0, 0.3}	Long periods of stability with shorter periods of substantial distance changes

Table 14.2 summarizes the used graph generators.

14.4 Experiment

For the test the JACSF - Java Ant Colony System Framework for TSP was used. It was described and made available to the research community by U. Chirico [3]. The modifications to the basic framework were marginal modifications and were introduced in order to make the test results easier to gather without modifying the core operation of an Ant Colony. Each graph generator was tested with at least 200 PTS that were generated in a way described in the Section 14.2.2.

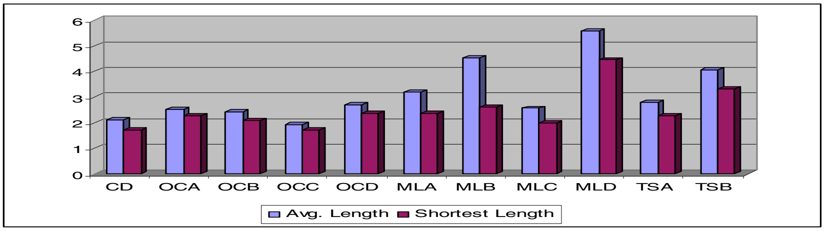


Fig. 14.1 Comparison of shortest and average route lengths

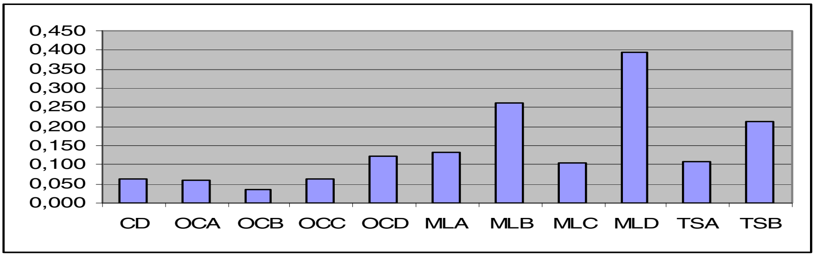


Fig. 14.2 Variation of the average route lengths

As indicated in the previous Chapter the graph generators work in a deterministic manner. This makes the comparison of results obtained for different PTS more reliable. To unsure such a mode of operation the random number generator used for generator operation was seeded with a constant value. On the other hand the operation of Ant Colony is random. Although the route selection is deterministic the initial positions of the ants are random. Therefore the shortest path lengths for the same test set could differ could differ. For that reason each PTS was used 5 times. The number of Ants was set to 50. It was the same as the number of graph nodes. The number of iterations was set to ItMax=700. In what follows the BPi denotes the length of the best path found up to the i-th iteration. When used without the index the BP refers to the length for the last iteration. It is used for static and relatively simple dynamic graphs. The performance analysis of ML and TS graph generators requires the running best path RunBP defined by the Formula 14.2.

$$RunBp = \left. \sum_{i=1}^{ItMax} BPi \middle/ ItMax \right. \tag{14.2}$$

The shortest and the running best paths for all parameter sets variation are presented on the Figures 14.1 and 14.2.

Note the high values of variation for the MLD and MLB what indicates that for this highly dynamic environments the proper selection of parameters is especially important.

14.4.1 Data Exploration

The Figure 14.3 depicts the RunBP with corresponding parameters for the TSB graph generator. The values were sorted according the acceding values of the RunBP. Figures for all other generators look much the same.

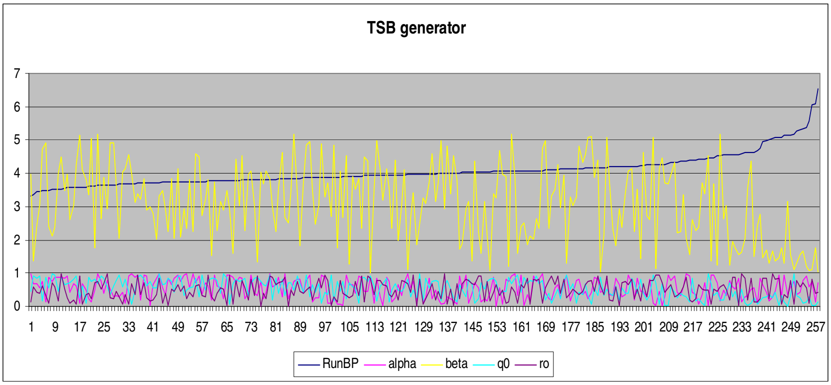


Fig. 14.3 Experiment results for the TSB graph generator

There are two conclusions that could be drawn from the figures. For the vast majority of parameter setting the length of the routes remain pretty much the same, the line representing the length remains near horizontal for much of the data. The route length diversity is higher for dynamic than static graph.

There is no easy detectable correlation between the route length and the parameter values. It is therefore necessary to apply statistical analysis in order to find correlation, if any, between RunBP and parameter values.

Table 14.3 Correlation between route lengths and parameter values

Code	Alpha	Beta	Q0	Ro
CD	***-0.327***	***-0.224***	***-0.502***	***0.140***
OCA	***-0.259***	***-0.482***	***-0.549***	0.098
OCB	***-0.384***	***-0.426***	***-0.424***	0.131
OCC	***-0.273***	***-0.456***	***-0.374***	***0.200***
OCD	***-0.385***	***-0.369***	***-0.463***	***0.154***
MLA	***-0.328***	***-0.310***	***-0.689***	***0.157***
MLB	***-0.175***	***-0.473***	***-0.476***	***0.045***
MLC	***-0.254***	***-0.413***	***-0.425***	0.025
MLD	***-0.145***	***-0.502***	***-0.485***	0.063
TSA	***-0.295***	***-0.352***	***-0.519***	0.083
TSB	-0.084	***-0.406***	***-0.472***	0.084

14.4.2 Correlation Analysis

The correlations between the BPi or RunBP and the individual parameters for all graph generators are presented in the Table 14.3. The statistically significant values for the confidence level = 0.05 are marked by bold digits.

Excluding the Ro parameter all others are clearly correlated with route lengths. The importance of β increases with the diversity level of the graph whereas for q0 the relationship is reverse. The correlation value for α is the strongest for a static graph and for all dynamic graphs its values decrease with the increase of graph changeability. The correlation of the lengths with the sum of parameters was also calculated. The results are presented in the Table 14.4.

Table 14.4 Correlation between route lengths and sum of parameter values

Code	Alpha+Beta	Alpha+Q0	Alpha+Ro	BetaQ0	BetaRo	QoRo
CD	***-0.395***	***-0.572***	***-0.125***	***-0.509***	***-0.059***	***-0.271***
OCA	***-0.528***	***-0.576***	***-0.112***	***-0.731***	***-0.299***	***-0.327***
OCB	***-0.583***	***-0.549***	***-0.180***	***-0.589***	***-0.218***	-0.193
OCC	***-0.535***	***-0.459***	***-0.059***	***-0.574***	-0.197	***-0.122***
OCD	***-0.535***	***-0.555***	***-0.162***	***-0.636***	-0.148	***-0.202***
MLA	***-0.456***	***-0.690***	-0.120	***-0.737***	***-0.108***	***-0.366***
MLB	***-0.490***	***-0.425***	-0.088	***-0.691***	***-0.302***	***-0.303***
MLC	***-0.496***	***-0.499***	***-0.165***	***-0.589***	***-0.273***	***-0.287***
MLD	***-0.486***	***-0.420***	***-0.054***	***-0.715***	***-0.328***	***-0.291***
TSA	***-0.450***	***-0.566***	***-0.136***	***-0.650***	***-0.177***	***-0.299***
TSB	***-0.337***	***-0.389***	-0.003	***-0.634***	***-0.214***	***-0.284***

Please note the outstanding correlation values in the Beta and Q0 column.

14.5 Conclusions

The experiments show the remarkable power of ACO to adopt itself to diverse sets of parameters values and produce acceptable results. For the comparison of different TSP Ant algorithms we require that the algorithms have optimal or near optimal parameter values. There are not many papers on the subject and experiments have shown that the values previously suggested do not lead to best results. The selection methods described in [4] treat all parameters in the same manner.

The process of parameters selection is time consuming and therefore it is desirable to lower the complexity of the task by concentrating upon parameters that have the greatest impact on the route length. The statistical analysis of the data from numerous experiments shows show that mostly influential are the β and Q0.

We are convinced that the proposed in the chapter graph generators cover a wider range of cases than the relatively simple graph modification methods proposed so far. Therefore we hope that they could be used by other researches in their study on the Dynamic TSP problem.

References

1. Dorigo, M.: Optimization, Learning and Natural Algorithms, PhD thesis, Politecnico di Milano, Italie (1992)
2. Dorigo, M., Stuetzle, T.: Ant Colony Optimization: Overview and Recent Advances, IRIDIA–Technical Report Series, Technical Report No. TR/IRIDIA/2009-013 (2009)
3. Chirico, U.: A Java Framework for Ant Colony Systems. In: Ants 2004: Forth International Workshop on Ant Colony Optimization and Swarm Intelligence, Brussels (2004)
4. Siemiński, A.: TSP/ACO Parameter Optimization; Information Systems Architecture and Technology; System Analysis Approach to the Design, Control and Decision Support, pp. 151–161. Oficyna Wydawnicza Politechniki Wrocławskiej (2011)
5. Psarafits, H.N.: Dynamic vehicle routing: Status and Prospects. National Technical Annals of Operations Research. University of Athens, Greece (1995)
6. Guntsch, M., Middendorf, M.: Pheromone modification strategies for ant algorithms applied to dynamic TSP. In: EvoWorkshops 2001: Appl. of Evol. Comput., pp. 213–222 (2001)
7. Guntsch, M., Middendorf, M.: A Population Based Approach for ACO. In: Cagnoni, S., Gottlieb, J., Hart, E., Middendorf, M., Raidl, G.R. (eds.) EvoIASP 2002, EvoWorkshops 2002, EvoSTIM 2002, EvoCOP 2002, and EvoPlan 2002. LNCS, vol. 2279, pp. 72–81. Springer, Heidelberg (2002)
8. Mavrovouniotis, M., Yang, S.: Ant Colony Optimization with Immigrants Schemes in Dynamic Environments. In: Schaefer, R., Cotta, C., Kołodziej, J., Rudolph, G. (eds.) PPSN XI. LNCS, vol. 6239, pp. 371–380. Springer, Heidelberg (2010)

Chapter 15
Tracking Changes in Database Schemas

Jakub Marciniak and Tadeusz Pankowski

Abstract. We discuss the problem of discovering changes in evolving XML schemas. Schema evolution is a natural, unavoidable phenomenon in contemporary data systems, that impacts both data transformation and query rewriting. We propose a rule-based algorithm that determines matched and unmatched schema elements thereby identifying changes in a schema under consideration. Additionally, we develop a method for computing edit distance in terms of some schema operations (insertion, deletion, renaming, and translocation). In result, we are able to obtain a set of operations which transform a given schema into the modified (target) form. The proposed algorithms have been fully implemented.

15.1 Introduction

It is natural and unavoidable that both data and schemas in contemporary systems continuously evolve, change and become more and more complex. The reason of this is that systems must be frequently adapted to real world changes and new functionalities must be introduced. The issue becomes considerably more complex when the schema under consideration is a result of integrating heterogeneous source schemas, especially when the collection of source schemas can dynamically change. In such environment the resulting (or target) schema is usually given in a form of XML schema, specified as DTD (*Document Type Definition*) or XSD (*XML Schema Definition*) [2, 10]. The flexibility of XML schema constructs admits the possibility

Jakub Marciniak
Faculty of Mathematics and Computer Science,
Adam Mickiewicz University, Poznan, Poland
e-mail: kubam@amu.edu.pl

Tadeusz Pankowski
Institute of Control and Information Engineering,
Poznań University of Technology, Poland
e-mail: tadeusz.pankowski@put.poznan.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 155–165.
springerlink.com

of much more complicated changes than those typically encountered in relational databases, making the XML schema evolution a difficult problem.

There are two important reasons why it is important to discover changes in the schema: (a) data transformation, and (b) query rewriting. The former is of special importance when data must be transformed from an instance of the base (source) schema into an instance of the modified (target) schema, and the latter when the queries passed against the base schema are to be rewritten into queries over the modified schema.

Discovering changes in database schema was first addressed with respect to relational and hierarchical database systems [9, 5]. In these approaches, the maintainer is responsible for explicitly describing the necessary transformation manually using a special purpose data translation language. Next, the problem was investigated for object-oriented database systems (e.g. [1]). The necessary transformation functions are then defined upon the changes made to the definitions of data types. However, manual comparison of complex schemas takes a lot of time and often some changes may be overlooked.

In this chapter we propose an automatic method that can be used to manage multiple XML schema structures which are frequently modified. The idea of the solution proposed in this chapter is as follows:

1. An algorithm for automatic discovering changes between a source schema and a target schema is proposed. We assume that the target schema arises from the source one in result of some structural modifications.
2. The outcome of the algorithm is a set of basic operations over the source schema which transform it to the target schema.
3. The algorithm is based on a set of hierarchically ordered schema matching rules which are successively applied to pairs of subtrees of the source and target schemas. Each rule is used to decide about the existence of matching relation between nodes.

15.2 Changes Discovering Process

Consider schema S_1 and a schema S_2 obtained from S_1 as the result of some structural modification. In this chapter we will call S_1 a source (base) schema and S_2 a target (modified) schema. The changes discovering process is a process in which we find how different (similar) are these two schemas and what changes have been made to the base schema. As the outcome we would like to acquire a list (preferably as small as possible) of operations that can be performed in order to transform schema S_1 to S_2. Using these operations we can easily define a mapping between these schemas and then transform any instance I_1 over schema S_1 to an instance I_2 that conforms to schema S_2 (Figure 15.1).

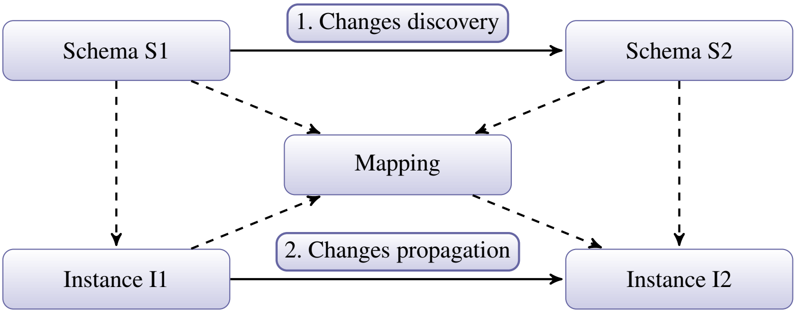


Fig. 15.1 Discovering changes process

15.2.1 Schemas Similarity

As the starting point for tracking changes we need a tree metric that can be used to find out differences between a base and a modified schema. There are several tree metrics published [3], however most of them are not feasible for XML schemas (labelled trees) and cannot be easily adapted to discover a mapping between schemas. The metric algorithms also depend much on the operations taken into consideration. For discovering changes between XML schemas, the following edit operations are useful:

- insertion (insertion of leaf nodes and insertion of node between existing nodes in hierarchy),
- deletion (deletion of leaf nodes and deletion of whole subtrees),
- renaming,
- translocation (detaching node or subtree from its parent and attaching it to another parent).

Unfortunately, while using all these operations, the problem of finding edit distance between trees is NP-complete. The best solutions we have found in the literature are the modified version of the Kleen algorithm [3] and the EditScript algorithm [4]. However both of these algorithms consider only limited set of edit operations and therefore for many cases they give very complicated results, e.g. subtree translocation is represented by multiple deletion and insertion operations. Both these algorithms assume also existence of initial stage of node matching (mapping).

As a running example we will consider a source XML schema S_1 depicted in Figure 15.2 and containing information about publications and their authors. There may be many publications and each publication can be written by many authors. We assume that this schema was modified. Some elements have been removed, some have been added and some have been moved to different place. The modified version of this schema is S_2 presented also in Figure 15.2. Even for these schema, containing only few nodes, it may take some time to point out all the differences. For a complex schema with hundred nodes or more, it will take hours to discover all the changes manually.

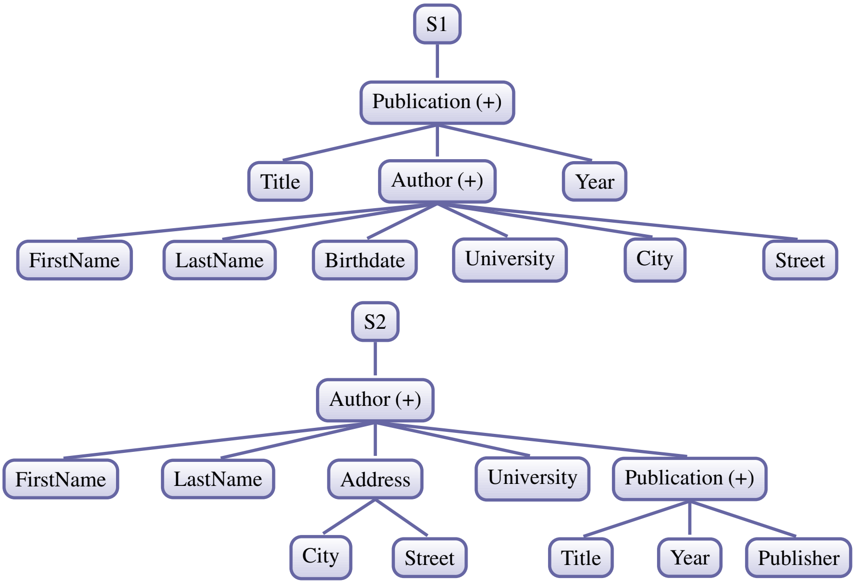


Fig. 15.2 A source (base) schema S_1 and the target (modified) schema S_2.

15.2.2 Rule Based Schema Matching

Now, we describe how automatically match nodes in schemas and calculate an edit distance between them. We propose simple and yet very flexible method that can be used to combine multiple criteria. We show that it gives very good practical results.

Existing matching algorithms are based upon some logic defined to find nodes that have some features in common e.g. label, parents, siblings or children. They give positive results for some cases, but for some they do not. Any attempt to correct the matching would require extensive knowledge of the algorithm and programming skills to change its implementation.

We suggest using set of rules (relations) describing common features of schema nodes. We are formalizing the rules using the following schema definition.

Definition 1 (Schema). A XML schema is a directed graph, $S =< \mathcal{N}, \mathcal{E}, \rho, l >$, where: (a) $\mathcal{N}$ is a finite set of nodes; (b) $\mathcal{E}$ is a finite set of directed edges between nodes, $(n_1 \rightarrow n_2) \in \mathcal{E}$; (c) ρ is the root label, $\rho \in \mathcal{N}$; (d) l is a function that assigns labels to nodes, $l : \mathcal{N} \rightarrow Str$.

We will also use $p(n)$ to denote the parent node of n, $(p(n), n) \in \mathcal{E}$, and $cn(n)$ to denote the set of children of n, $cn(n) = \{c \mid (n, c) \in \mathcal{E}\}$, and $\simeq$ to denote the equivalence relation on node sets: $N_1 \simeq N_2 \iff \forall_{n_1 \in N_1} \exists_{n_2 \in N_2} l(n_1) = l(n_2) \wedge \forall_{n_2 \in N_2} \exists_{n_1 \in N_1} l(n_1) = l(n_2)$, i.e. N_1 and N_2 are equivalent if they have the same number of nodes with the same labels.

Rules of matching: Let S_1 be a base schema and S_2 be its modified version and $\rho_1, n_1 \in S_1$ and $\rho_2, n_2 \in S_2$. We use $\sim_i$ to denote the $i-$th node matching relation, i.e. the matching relation determined by the i-th rule defined below. By $\sim$ we will denote that some matching relation holds between nodes, i.e. $n_1 \sim n_2 \leftrightarrow \exists_i n_1 \sim_i n_2$. Matching between sets, $N_1 \sim_i N_2$, holds when each node from one set has exactly one matched node in the other one.

1. ρ_1 is a root node and ρ_2 is a root node $\Rightarrow r_1 \sim_1 r_2$
2. $p(n_1) \sim p(n_2) \wedge l(n_1) = l(n_2) \wedge cn(n_1) \simeq cn(n_2) \Rightarrow n_1 \sim_2 n_2$
3. $l(n_1) = l(n_2) \wedge cn(n_1) \simeq cn(n_2) \Rightarrow n_1 \sim_3 n_2$
4. $p(n_1) \sim p(n_2) \wedge l(n_1) = l(n_2) \Rightarrow n_1 \sim_4 n_2$
5. $p(n_1) \sim p(n_2) \wedge cn(n_1) \simeq cn(n_2) \Rightarrow n_1 \sim_5 n_2$
6. $p(n_1) \sim p(n_2) \wedge cn(n_1) \sim cn(n_2) \Rightarrow n_1 \sim_6 n_2$
7. $l(n_1) = l(n_2) \Rightarrow n_1 \sim_7 n_2$
8. $cn(n_1) \sim cn(n_2) \wedge cn(n_1) \neq \emptyset \Rightarrow n_1 \sim_8 n_2$

Presented rules have been chosen experimentally to work well in most cases. We assume that rules should be ordered accordingly to the probability that nodes really match each other. Of course for specific cases the rules should be tuned a little bit to obtain better results. Adjusting them can be done quite easily and does not require any interference within the algorithm.

The matching algorithm checks the rules in the order they were given. We take the first rule and try to apply it to any pair of nodes. If a rule condition is true then we have found a match and we don't check these nodes again, but we start over from rule number one (or rule number two as the first rule is just a starting point). If i-th rule could not be applied at all, then rule $i+1$ is being tested. This way we should maximize the probability of correct matching (as the rules order should be based upon matching probability). The matching process can be then performed by the following algorithm (R is a rule set).

Algorithm 1. Matching schema nodes

```
function MATCHNODES($\mathscr{N}_1, \mathscr{N}_2, R$)
    $\mathscr{M} \leftarrow \emptyset$                                  ▷ Matched nodes
    $i \leftarrow 1$                                  ▷ Rule number
    rulesLoop:
    while $i \leq |R|$ do
        for all $n_1 \in \mathscr{N}_1$ do
            for all $n_2 \in \mathscr{N}_2$ do
                if $n_1 \sim_i n_2$ then                  ▷ Nodes matched by rule i
                    $\mathscr{M} \leftarrow \mathscr{M} \cup (n_1, n_2)$
                    $\mathscr{N}_1 \leftarrow \mathscr{N}_1 \setminus \{n_1\}$
                    $\mathscr{N}_2 \leftarrow \mathscr{N}_2 \setminus \{n_2\}$
                    $i \leftarrow 1$                  ▷ Start over from first rule
                    continue rulesLoop
        $i \leftarrow i + 1$                              ▷ Check next rule
    return $\mathscr{M}$                              ▷ Return matches
```

As a result, for our running example, we have the following matching between schemas

- $/S1 \sim /S2$
- $/S1/Publication \sim /S2/Author/Publication$
- $/S1/Publication/Author \sim /S2/Author$
- $/S1/Author/Publication/Title \sim /S2/Publication/Title$
- $/S1/Publication/Year \sim /S2/Author/Publication/Year$
- $/S1/Author/LastName \sim /S2/Publication/Author/LastName$
- $/S1/Author/FirstName \sim /S2/Publication/Author/FirstName$
- etc...

The nodes that could not be matched are /S1/Publication/Author/Birthdate, /S2/Author/Address, /S2/Author/Publication/Publisher.

15.2.3 Edit Distance

In this section we describe how an edit distance between schemas can be found. Because our algorithm gives answer for the NP-complete problem in polynomial time it may not give the best possible result. However the tests we have evaluated show that it is quite good approximation.

We use an adaptation of well known Levenshtein distance as a base metric between two schema nodes, and using it recurrently we can find the distance between schema trees. Levenshtein distance is a metric typically used for measuring the difference between two words (sentences). Therefore, it normally compares characters in the words. For our application it compares node labels.

The following algorithms analyze children of a single node from the base schema and children of a single node from the modified schema. Consider example from Figure 15.2 and focus on the node *Author*. We compare its child nodes *(FirstName, LastName, Birthdate, University, City, Street)* with nodes *(FirstName, LastName, Address, University, Publication)*. We can see that nodes *Address* and *Publication* have been added, while *Birthdate*, *City* and *Street* have been removed.

The original algorithm for calculating Levenshtein distance uses basic operations (insertion, renaming, deletion) and checks all possible sequences of these operations to find distance between two input character chains. We restrict this algorithm to find only insertion and deletion operations and we combine the outcome with the results of Algorithm 1. This way we can discover all basic operation types (insertion, deletion, renaming, translocation).

As the result from the algorithm above we obtain following distance matrix. Integer values represent cost of transforming one node set to the other. Transition in column corresponds to adding a node, while transition in row corresponds to deleting a node. Diagonal transition is only permissible if nodes in row and column have same name and no change is needed.

Example 1. Levenshtein distance matrix

	$\perp$	FirstName	LastName	Birthdate	University	City	Street
$\perp$	**0,00**	1,00	2,00	3,00	4,00	5,00	6,00
FirstName	1,00	**0,00**	1,00	2,00	3,00	4,00	5,00
LastName	2,00	1,00	**0,00**	1,00	2,00	3,00	4,00
Address	3,00	2,00	**1,00**	**2,00**	3,00	4,00	5,00
University	4,00	3,00	2,00	3,00	**2,00**	3,00	4,00
Publication	5,00	4,00	3,00	4,00	**3,00**	**4,00**	**5,00**

Algorithm 2. Levenshtein distance

```
function LEVENSHTEINDISTANCEMATRIX($\mathcal{N}_1, \mathcal{N}_2$)
    $m \leftarrow |\mathcal{N}_1|, n \leftarrow |\mathcal{N}_2|$
    $d \leftarrow [0..m, 0..n]$                              ▷ Distance matrix
    for $i \leftarrow 0$ to $m$ do                            ▷ Initialize columns
        $d[i,0] \leftarrow i$
    for $j \leftarrow 0$ to $n$ do                            ▷ Initialize rows
        $d[0,j] \leftarrow j$
    for $j \leftarrow 0$ to $n$ do
        for $i \leftarrow 0$ to $m$ do
            $n_i \leftarrow \mathcal{N}_1^i$                  ▷ i-th node from $N_1$
            $n_j \leftarrow \mathcal{N}_2^j$                  ▷ j-th node from $N_2$
            if $l(n_i) = l(n_j)$ then                         ▷ Same names,
                $d[i,j] \leftarrow d[i-1, j-1]$               ▷ no operation required
            else
                $d[i,j] \leftarrow \min($
                  $d[i-1, j] + 1,$                            ▷ Delete operation
                  $d[i, j-1] + 1)$                            ▷ Insert operation
    return $d$                                                ▷ Return distance matrix
```

15.2.4 Transforming Operations

Using distance matrix we can easily find the smallest possible set of basic operations transforming one sequence to another. In the matrix d each path from $d[0,0]$ to $d[m,n]$ represents some operations that can transform N_1 to N_2, e.g. $(0,0) \rightarrow (0,1) \rightarrow ... \rightarrow (0,n) \rightarrow (1,n) \rightarrow ... \rightarrow (m,n)$ corresponds to removing all source labels and adding all target labels. However that is not the path we are searching for. We already know the size of minimal operation set because it is equal to $d[m,n]$. Therefore we have to find a path that has most transitions of type $(i,j) \rightarrow (i+1, j+1)$ (for nodes that have the same names). Starting from $d[m,n]$ we find a path to $d[0,0]$ containing elements with smallest values in the matrix (Algorithm 3).

Algorithm 3. Basic transforming operations

```
function FINDBASICOPERATIONS($d[0..m,0..n],\mathcal{N}_1,\mathcal{N}_2$)
    $O_1 \leftarrow \emptyset$
    $i \leftarrow m$, $j \leftarrow n$
    while $i > 0 \wedge j > 0$ do
        if $d[i-1][j-1] = d[i][j]$ then
            $i \leftarrow i-1$, $j \leftarrow j-1$                                    ▷ Same node labels
        else if $d[i][j-1] < d[i-1][j]$ then
            $O_1 \leftarrow O_1 \cup (INSERT, \mathcal{N}_2^{j-1})$, $j \leftarrow j-1$          ▷ Insert ($j$-th−1) node
        else
            $O_1 \leftarrow O_1 \cup (DELETE, \mathcal{N}_1^{i-1})$, $i \leftarrow i-1$          ▷ Delete ($i$-th−1) node
    while $i > 0$ do                                                 ▷ Delete remaining nodes
        $O_1 \leftarrow O_1 \cup (DELETE, \mathcal{N}_1^{i-1})$, $i \leftarrow i-1$
    while $j > 0$ do                                                 ▷ Insert remaining nodes
        $O_1 \leftarrow O_1 \cup (INSERT, \mathcal{N}_2^{j-1})$, $j \leftarrow j-1$
    return $O_1$                                                     ▷ Return basic operation
```

Definition 2 (Operation). A transforming operation for schema $S =< \mathcal{N}, \mathcal{E}, \rho, l >$ is a pair $op =< \mathcal{T}, n >$, where: (a) $\mathcal{T}$ is a type of operation (INSERT, DELETE, MOVE, RENAME); (b) n is a node from schema, $n \in \mathcal{N}$.

For our example path $(0,0) \rightarrow (1,1) \rightarrow (2,2) \rightarrow (2,3) \rightarrow (3,3) \rightarrow (4,4) \rightarrow (4,5) \rightarrow (5,5) \rightarrow (5,6)$ represents the minimal operation set O={INSERT(Address), REMOVE(Birthdate), INSERT(Publication), REMOVE(City), REMOVE(Street)}. As you can see cardinality of this set is equal to $d[m,n]$ in the Levenshtein distance matrix.

The obtained set of transforming operations contains only information about inserted and deleted nodes. Combining this information with the results from Algorithm 1, we discover additional transforming operations (renaming, translocation). Especially, when an inserted node has a matched node, it must have been moved to a different place in the schema. For the same reason we ignore deletions for matched nodes as they will be included by some translocation operation (Algorithm 4).

Then we execute described steps recurrently for all nodes in modified schema. Transforming operation, found in the process, must be applied to base schema. As the final result we obtain a list of operations that can be used to transform base schema to its modified version. The size of this list is the edit distance between schemas.(Algorithm 5)

For our example the discovered transforming operation are:

- $RENAME(/S1 \rightarrow /S2)$
- $MOVE(/S1/Publication/Author \rightarrow /S2/Author)$
- $INSERT(/S2/Author/Address)$
- $DELETE(/S1/Publication/Author/Birthdate)$
- $MOVE(/S1/Publication \rightarrow /S2/Author/Publication)$

- $MOVE(/S1/Publication/Author/City \rightarrow /S2/Author/Address/City)$
- $MOVE(/S1/Publication/Author/Street \rightarrow /S2/Author/Address/Street)$
- $INSERT(/S2/Author/Publication/Publisher)$

Algorithm 4. Transforming operations

```
function FINDTRANSFORMINGOPERATIONS($O_1, \mathscr{M}$)
    $O_2 \leftarrow \emptyset$
    for all $op \in O_1$ do
        $\mathscr{T} \leftarrow \mathscr{T} | \exists_n (\mathscr{T}, n) = op$          ▷ Operation type
        $n_1 \leftarrow n | (\mathscr{T}, n) = op$          ▷ Node from operation
        $n_2 \leftarrow n | (n_1, n) \in \mathscr{M}$          ▷ Matched node
        if $n_2 = \perp$ then          ▷ No match found
            $O_2 \leftarrow O_2 \cup op$
        else if $\mathscr{T} = INSERT$ then          ▷ Match found
            if $l(n_1) = l(n_2)$ then
                $O_2 \leftarrow O_2 \cup \{(MOVE, n_2)\}$          ▷ Add translocation operation
            else
                $O_2 \leftarrow O_2 \cup \{(RENAME, n_2)\}$          ▷ Add renaming operation
    return $O_2$          ▷ Return transforming operation
```

Algorithm 5. Tree edit operations

```
function TREEEDITOPERATIONS($n_1, n_2, \mathscr{M}$)
    $O \leftarrow \emptyset$          ▷ Set with all operations
    $C_1 \leftarrow cn(n_1), C_2 \leftarrow cn(n_2)$          ▷ Get children sets
    $d \leftarrow$ LEVENSHTEINDISTANCEMATRIX($C_1, C_2$)          ▷ Create distance matrix
    $O_1 \leftarrow$ FINDBASICOPERATIONS($d, C_1, C_2$)          ▷ Find basic operations
    $O_2 \leftarrow$ FINDTRANSFORMINGOPERATIONS($O_1, \mathscr{M}$)          ▷ Find all operations
    $C_1 \leftarrow$ PERFORMOPERATIONS($n_1, O_2$)          ▷ Perform operations on base schema
    for all $c_2 \in C_2$ do
        $c_1 \leftarrow c | l(c) = l(c_2)$          ▷ Find node with same name
        $O \leftarrow O \cup$ TREEEDITOPERATIONS($c_1, c_2, \mathscr{M}$)          ▷ Recurrent call
    return $O$          ▷ Return operations
function SCHEMAEDITDISTANCE($S_1 =< \mathscr{N}_1, \mathscr{E}_1, \rho_1, l_1 >, S_2 =< \mathscr{N}_2, \mathscr{E}_2, \rho_2, l_2 >$)
    $\mathscr{M} \leftarrow$ MATCHNODES($\mathscr{N}_1, \mathscr{N}_2, R$)          ▷ R is a rules set
    $O \leftarrow$ TREEEDITOPERATIONS($\rho_1, \rho_2, \mathscr{M}$)          ▷ Find operations starting from roots
    return $|O|$          ▷ Edit distance
```

As we can see our algorithms produces satisfying results for our running example. The example was chosen to show that our method works even for quite complex structural modifications. Typically subsequent schema versions are not that different from each other and can be easily discovered by presented algorithms.

15.3 Conclusions

We have discussed method for tracking changes in dynamic XML database systems. It can be used to compare two different versions of the same schema or to constantly monitor existing schemas. It provides enough information for making adequate decisions e.g. modifications in existing software, changing queries, modification in data export/import scripts.

Together with the schema extraction [8, 6] methods it can be even used to analyze and validate incoming data, automatically create a schema for them or even convert them to desired format. We have prepared methods to automatically propagate discovered changes to instance documents, but due to space limitations we can't present it in this chapter.

We believe described methods may have many applications and can save users a lot of effort.

The discussed algorithms are implemented within XTR system, which is available at `http://www.xtr.sf.net` [7]. Application is free and open source with the following main features:

- XML Transformation
- XML, XSD Summarization
- XML Merging
- XSD Schema extraction - xml2xsd
- XSD Comparison
- XML Validation (Syntax validation, Validation with XML Schema - XSD)
- XML Edition - with auto formatting and syntax highlighting
- XSD visualization (using tex with tikz library)

References

1. Banerjee, J., Kim, W., Kim, H.J., Korth, H.F.: Semantics and implementation of schema evolution in object-oriented databases. In: SIGMOD Conference, pp. 311–322. ACM Press (1987)
2. Bex, G.J., Neven, F., den Bussche, J.V.: DTDs versus XML Schema: A Practical Study. In: WebDB, pp. 79–84 (2004)
3. Bille, P.: A survey on tree edit distance and related problems. Theor. Comput. Sci. 337(1-3), 217–239 (2005)
4. Chawathe, S.S., Rajaraman, A., Garcia-Molina, H., Widom, J.: Change detection in hierarchically structured information. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 493–504 (1996)
5. Lerner, B.S.: A model for compound type changes encountered in schema evolution. ACM Trans. Database Syst. 25(1), 83–127 (2000)
6. Marciniak, J.: XML Schema and Data Summarization. In: Rutkowski, L., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2010. LNCS(LNAI), vol. 6114, pp. 556–565. Springer, Heidelberg (2010)

7. Marciniak, J., Pankowski, T.: Automatic xml data transformation and merging. Zeszyty Naukowe Wydzialu ETI Politechniki Gdańskiej. Technologie Informacyjne 16, 231–236 (2008)
8. Martens, W., Neven, F., Schwentick, T.: Simple off the shelf abstractions for XML schema. SIGMOD Record 36(3), 15–22 (2007)
9. Navathe, S.B.: Schema analysis for database restructuring. ACM Trans. Database Syst. 5(2), 157–184 (1980)
10. W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes (2009), www.w3.org/TR/xmlschema11-2

Part III
Information Systems Applications

Chapter 16
Smart Communication Adviser for Remote Users

Miroslav Behan and Ondrej Krejcar

Abstract. Today's various innovations in communication technology environment lay down foundations for future smart solutions through sectors. The abundance of possibilities in personal communication led us to simplify the process from user point of view. Therefore, we propose the concept for convenient and environmentally smart based applications which are focused on usability and clarity of information.

16.1 Introduction

The infinite possibilities of nowadays communication which are developing further more thanks to miniaturization of processing power and increasing throughput of mobile networks led us to the idea that if there would be an independent mobile application which could easily advise users in today's confusing bundle of services and if could save a budget spent on communication.

In last decade mobile technology due to its penetration crosses the global world starting the competition between quality, cost and usability of communication services. The last one criterion seems to be arising of importance in future. We can recognize competition among leaders of mobile device's platforms such as Android, iOS, Windows Mobile, PalmOS, Bada, SymbianOS or MeeGo. We consider just first three players in future market due to their current influence with information synergy power. Nokia lost leading position on mobile market caused by ignoring open development power and by over flooded market with too many useless devices.

Miroslav Behan · Ondrej Krejcar
University of Hradec Kralove, FIM, Department of Information Technologies
Rokitanskeho 62, Hradec Kralove, 500 03, Czech Republic
e-mail: miroslav.behan@uhk.cz, ondrej.krejcar@asjournal.eu

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 169–178.
springerlink.com

The time of closed development is definitely over and the open development we acknowledge as a mainstream for mobile device evolution progress which comes out from creativity, social networked and intellectual power hided in great number of many individuals. Today most people using mobile devices primarily for daily communication. However by changing establishment of current mobile network provider's profits which is based on voice and SMS to provide better customer usability, we see more and more only internet online connected devices use cases where are combined all possibilities of smart device features and where one single mobile device would step up all personal common needs [14].

We realize that the highest penetration of mobile device world widely are cellular phones where access to internet is limited by mobile provider for example by wireless application protocol (WAP) and installed applications are dictated by device's vendors or reduced with minimal hardware device interface knowledge. But we recognize that the trend is getting over with an increasing ratio of smart phones where access to the internet is standardized over GPRS, EDGE or HSDPA [1, 17]. The installment of application is independent process corresponding with user will and knowledge [16]. Last trend that outlines mobile device is wireless local area network (WLAN). The penetration of WLAN routers rapidly increased in last decade and also number of smart phones with WLAN capability growth in last years. Therefore, home, workplace and public environment create beneficial usability cases with WLAN networks which are connected to the internet [5]. For example in home environment user would have fast and free communication reality without mobile provider's internet needs [6-8].

16.2 Problem Definition

A minimum knowledge about communication possibility, which is acknowledged as trend and which decreases power to change mobile device usability further to use only internet network access, we define as a society problem [9]. To change nowadays establishment of mobile network providers we announce smart communication advisory based on environment and users behaviour where natural competition with quality and cost of services is the case [10, 11, 18]. Following fundamentals of communication principals are basically the same.

- **Text** – The written type of communication benefits with the accuracy information. Therefore, the conversation history and full text search can be provided when required.
- **Voice –** The spoken type of communication leads to fast acquirement in mind correlation of actors by intonation and stress in their voices.
- **Visualization** – This part we considered as supportive to the previous ones. Although visual perception of human in total amount of information is major source, for some circumstances the anonymisation of environment where mimics and background scene picture would lead to distortion in communication process even if we indicate non-visual and face-face communication as equivalent in terms of quality of life [3] (QoL) perception.

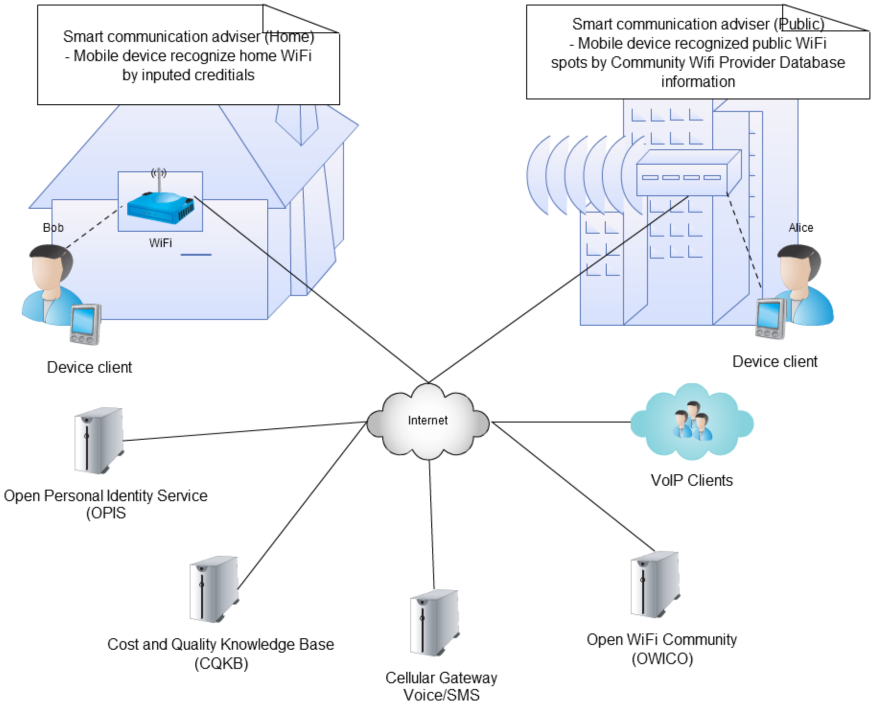

Smart communication adviser (Home)
- Mobile device recognize home WiFi
by inputed creditials
Smart communication adviser (Public)
- Mobile device recognized public WiFi
spots by Community Wifi Provider Database
information
Bob
WiFi
Alice
Device client
Device client
Internet
VoIP Clients
Open Personal Identity Service
(OPIS
Cost and Quality Knowledge Base
(CQKB)
Cellular Gateway
Voice/SMS
Open WiFi Community
(OWICO)

Fig. 16.1 Smart Communication Adviser (SCA) scenario [4]

There are also different relation types of communication such as one-one, one-many or many-many in bi-direction relations and also where we would consider active (synchronized) or passive (asynchronous) interactions between actors [12, 13]. The next perspective is about networks and current possibilities.

Networks for mobile devices are the main issue in the mobile communication and how mobile devices could access to network. One of the well-known standards used by mobile providers is a global system for mobile communication (GSM). Another standard that we consider is a wireless local area network (Wi-Fi). In the first case we consider the quality of service (QoS) in short main ones as second generation 2G (GPRS), 2,5G (EDGE), 3G (UTMS) or 4G (LTE). The network land coverage is analogically decreasing with an increasing generation level. In second case is interesting standard IEEE 802.11e which is supporting QoS of voice over IP (VoIP).

While we overviewed the technology aspects the extremely influenced communication behaviour are protocols and standards. The message based communication is well-known standard for mobile devices and as current mainstream we can announce short message service (SMS) where are technical restrictions for example length of transmitted message and lack of user’s status acknowledgment. The maximum length of message is defined up to 160 characters coded with 7bits, 140 characters by 8bits or 70 chars with 16bits. The notification of receiver's status or rather status of message are limited to short message service centre (SMSC) which is used for a correct message delivery. Therefore we considered SMS standard of communication as obsolete and this type we considered only as necessary bridge between the past and future approach to a message communication.

The influences of social media are resonating in current development possibilities more and more. Reasonable purpose of such user's behaviour we would realize in social environment based on daily human needs where the portion of social information bundle is as required as a food. The social information of interpersonal circle based on relationships are with its subjectivity and message importance overpassing other information which are based on a global knowledge without non-relationships interactivity. For this reason the natural usability of social instant messaging increases and we recognize this message communication as a future mainstream.

At last we consider a voice as a main type of communication. The bi-direction speech force actors to active expression type of process. The advantage to point out is the fast knowledge description but in comparison with message based communication the absence of quality of an informational history communication externalization we consider as a disadvantage. The technical aspect of voice exchange is real time network latency needs. Therefore we considered in concept real time protocol (RTP) for data exchange with low latency [2] with combination of session initial protocol (SIP) as communication control flow. All client-server-client cases consider latency and speed of network for correct recommendation in terms of quality available services.

The following table (see Table 16.1) tested server/client technologies describes android platform and appropriate communication message based technologies.

Table 16.1 Local Communication Methods

Local round trip measurement of Request/Response with object persistence on server side tested with client Android mobile device (ZTE Blade)	
Communication Method	Technology/Latency
AppServer/DB Engine/HTTP-POST	Appengine/Objectify(JPA)/250ms-400ms (avg. 300ms)
AppServer/Remote Cloud DB/HTTP-POST	Appengine/MySQL/450ms-600ms (avg. 500ms)
AppServer/Local DB/HTTP-POST	Tomcat7/ObjectDB/350ms-450ms (avg. 370ms)
JVM Server/DB Engine/TCP/Socket-Object Serialization	Socket Server/ObjectDB/20ms-30ms (avg. 23ms)

16.3 Solution Design

We present smart communication platform only as a concept in high level description. We consider the context of description for basic scenarios as representation of the most common current cases and we focus on a communication process itself (see Figure 16.2).

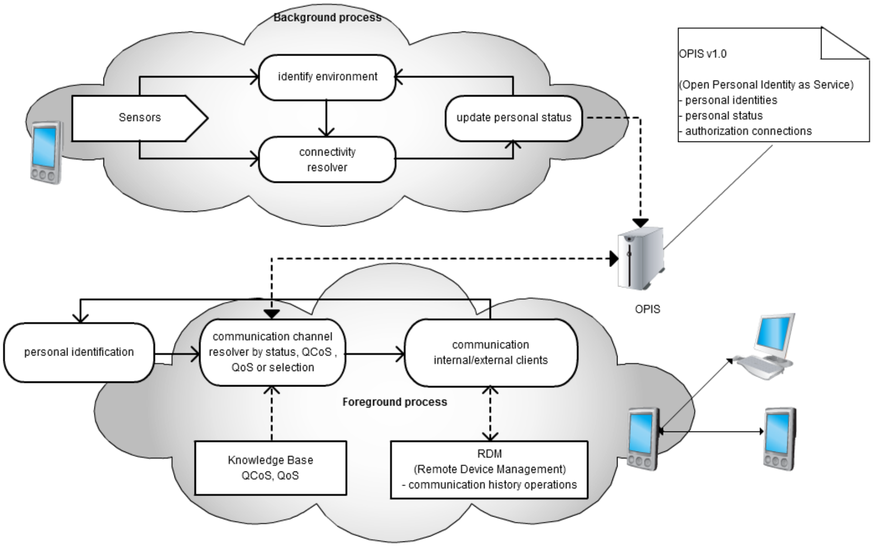

Background process
identify environment
Sensors
update personal status
connectivity resolver
OPIS v1.0
(Open Personal Identity as Service)
- personal identities
- personal status
- authorization connections
OPIS
personal identification
communication channel resolver by status, QCoS , QoS or selection
communication internal/external clients
Foreground process
Knowledge Base QCoS, QoS
RDM (Remote Device Management) - communication history operations

Fig. 16.2 Schema of communication process [4]

The home environment where access to network is provided by home WLAN access point (AP) personal mobile device is able to recognize actual environment of user location by inputted wireless network credentials, docking station or GPS location. Approach for environment recognition is defined by user's input. In consideration we recommend the lowest battery capacity with minimal impact on sensor checking. In some circumstances the recognition only by Wi-Fi connection would be sufficient for smart behaviour.

We divided concept into three independent parts which we consider as mandatory for success smart communication advisory behaviour. The first one we would define as an independent public personal profiles service (IPPP) which is basically gathering information from different resources as social networks, messaging servers and other external authorized inputs. This service has to be responsible for personal location and connectivity status based on availability of services and current environment. Beneficial would be personal identification related to more than one communication device.

Second mandatory part of the concept is an independent information service where cost and quality descriptions are located as a knowledge base for mobile application. We would call that kind of service as cost and quality knowledge base (CQKB) which will increase in precision in time.

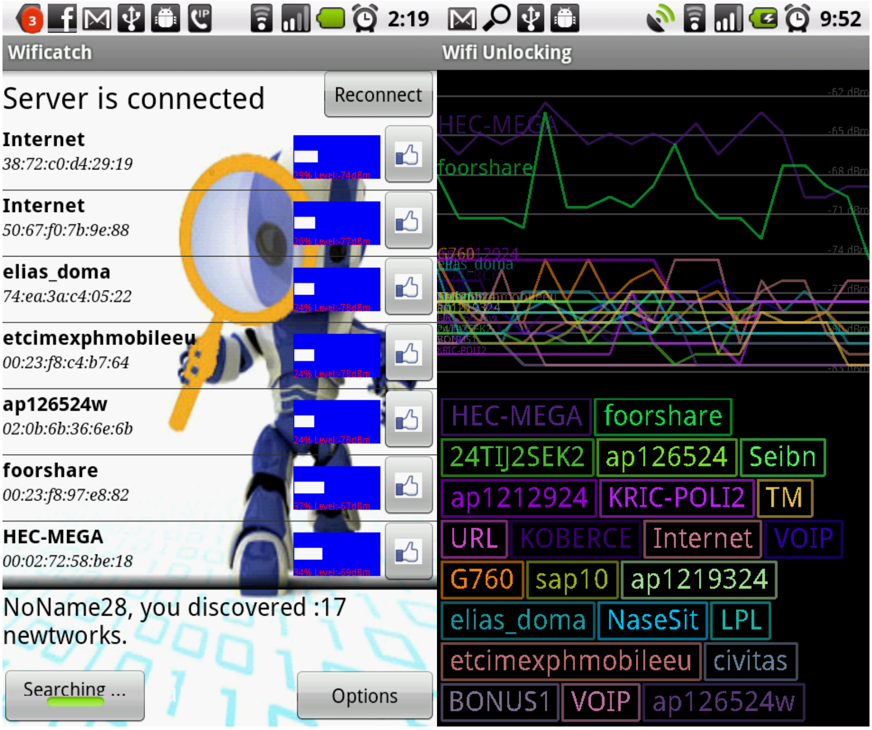
The next required part of the concept has to be the interface to cellular mobile devices which are marked in the schema as cellular gateway which is necessary to be there as an extensional bridge in terms of concept usability. And the last mandatory part that we consider is mobile device application as iPhone, Android and Windows-mobile platform based on client which informs user about quality and cost of service (QCoS) according to a location and environment. The supportive part of concept would be open Wi-Fi community (OWICO) where independent Wi-Fi providers could by micro payments propose fair low cost internet connectivity. For better comprehension next chapter describes communication process (see Figure 16.1).

We divided communication process into several parts. As the first one we consider personal identification with who is desired to communicate. Second would be about to establish connection between communication's actors in online or offline mode. That influence selection of communication services in terms of required quality, cost and availability. After selection or default per-defined configuration is communication itself established by internal or external service. Optionally history of interactions is stored as cloud service for future usage on different devices accessible over web client or this mobile client. The whole process ends by user’s action or service failure.

In the consideration of knowledge based on cost of service (CoS) as data-set acquired from mobile network provider's price lists. These sources are available on provider’s websites, which would be processed automatically by website parser or input manually by human operator into system. Other possible way would be by agreement with provider to supply information by external extraction over for example XML data format. The knowledge of quality of service (QoS) is based on empirical data gathered from the mobile device applications where specified circumstances have impact on precise network measurement.

16.4 Application Prototype

As part of the concept we designed android mobile application available on Google play market which is resolving network performance and which is gathering collected information about Wi-Fi networks and send them to remote community server (OWICO). For comprehensive interpretation we attach screenshot of application in following Figure 16.3 where are overviewed discovered Wi-Fi networks with signal, speed and latency criteria marked by current location. As the location provider is used GPS sensor or Internet connection knowledge, but in concept based on Wi-Fi community providers also Wi-Fi itself would be a provider of location as offline service.



2:19

Wificatch

Server is connected

Internet
38:72:c0:d4:29:19
29% Level:-74dBm

Internet
50:67:f0:7b:9e:88
26% Level:-77dBm

elias_doma
74:ea:3a:c4:05:22
24% Level:-78dBm

etcimexphmobileeu
00:23:f8:c4:b7:64
24% Level:-78dBm

ap126524w
02:0b:6b:36:6e:6b
24% Level:-78dBm

foorshare
00:23:f8:97:e8:82
37% Level:-67dBm

HEC-MEGA
00:02:72:58:be:18
34% Level:-69dBm

NoName28, you discovered :17 newtworks.

Fig. 16.3 Android Wi-Fi application

The collection of measured data delivered to server (OWICO) is related to Wi-Fi network quality and location. We expressed a detailed description of gathered information in Figure 16.4. For more precise results for building global knowledge based system, the independent user community would start up and is encouraged by free of charge access to internet by Wi-Fi networks at high frequent locations. As possible enhancement of prototype we considered the environmental behaviour for instance when mobile device stops moving it triggers background processes and searches the best QCoS and reassigns communication clients or changes communication status for allowed viewers and runs predefined tasks in recognized environment.

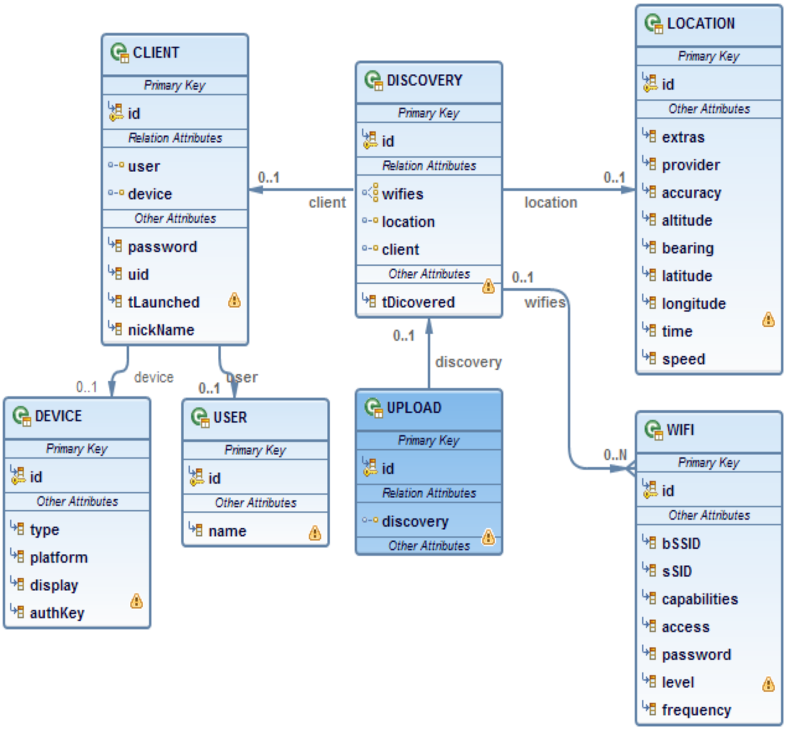


Fig. 16.4 Entity Relation Diagram

16.5 Conclusions

The goal of the article is in a society contribution with the concept of communication adviser application to provide a cost-effective knowledge. We highlighted mandatory definitions of exploring area in conceptual design and mind mapping of communication advisory problematic. The usage of application on daily bases would lead users to a cost-effective behaviour and global optimization of communication networks and to solve current problem with a lack of knowledge from mobile network providers in terms of providing communication service costs. The result of a mind experiment produced by designing technological concept of smart communication adviser leads into the creation of independent required services and encouraged to develop a real communication client which supports smart environment concepts where behaviour of users and devices would depend on a global knowledge. The challenge is about consideration of home, work and public environment as well as network capabilities as 2G, 3G, 4G and Wi-Fi. The aspect like active or passive communication type dramatically influence user's behaviour, QCoS and technological requirements. The concept supposed to be a pattern for development cross-platform application which provides advisory in inter-personal communication and inspire of cohesion services possibilities.

Acknowledgment. The work and the contribution were partially supported by the project (1) ESF funded project “INDOP – Innovation and Support of the Doctoral Study Program”, identification number CZ.1.07/2.2.00/28.0327; (2) “SMEW – Smart Environments at Workplaces”, the Grant Agency of the Czech Republic, GACR P403/10/1310; (3) specific research project "Smart Solutions in Ambient Intelligent Environments", University of Hradec Kralove under the project SP/2012/6.

References

1. Burakowski, W., Beben, A., Tarasiuk, H., Sliwinski, J., Janowski, R., Batalla, J.M., Krawiec, P.: Provision of end-to-end QoS in heterogeneous multi-domain networks. Annales des Telecommunications - Annals of Telecommunications 63(11-12), 559–577 (2008), doi:10.1007/s12243-008-0060-3
2. Liu, L.S., Zimmermann, R.: Measured end-to-end delay. Multimedia Systems 11(6), 497–512 (2005)
3. Lee, P.S.N., Leung, L., Lo, V., Xiong, C., Wu, T.: Social Indicators Research. Internet Communication Versus Face-to-face Interaction in Quality of Life 100(3), 375–383 (2011)
4. Online Diagram Designer, Creately. com, `https://creately.com/app/#`
5. Krawiec, J., Penhaker, M., Krejcar, O., Novak, V., Bridzik, R.: System for Storage and Exchange of Electrophysiological Data. In: Proceedings of 5th International Conference on Systems, ICONS 2010, April 11-16, pp. 88–91. IEEE Conference Publishing Services, Menuires (2010) ISBN 978-0-7695-3980-5, doi:10.1109/ICONS.2010.23
6. Korpas, D., Halek, J.: Pulse wave variability within two short-term measurements. Biomedical Papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia 150(2), 339–344 (2006) ISSN: 12138118
7. Kasik, V., Penhaker, M., Novák, V., Bridzik, R., Krawiec, J.: User Interactive Biomedical Data Web Services Application. In: Yonazi, J.J., Sedoyeka, E., Ariwa, E., El-Qawasmeh, E. (eds.) ICeND 2011. CCIS, vol. 171, pp. 223–237. Springer, Heidelberg (2011)
8. Brida, P., Machaj, J., Benikovsky, J., Duha, J.: An Experimental Evaluation of AGA Algorithm for RSS Positioning in GSM Networks. Elektronika ir Elektrotechnika 8(104), 113–118 (2010) ISSN 1392-1215
9. Chilamkurti, N., Zeadally, S., Jamalipour, S., Das, S.K.: Enabling Wireless Technologies for Green Pervasive Computing. EURASIP Journal on Wireless Communications and Networking, Article ID 230912, 2 pages (2009)
10. Chilamkurti, N., Zeadally, S., Mentiplay, F.: Green Networking for Major Components of Information Communication Technology Systems. EURASIP Journal on Wireless Communications and Networking, Article ID 656785, 7 pages (2009)
11. Liou, C.-Y., Cheng, W.-C.: Manifold Construction by Local Neighborhood Preservation. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 683–692. Springer, Heidelberg (2008)
12. Juszczyszyn, K., Nguyen, N.T., Kolaczek, G., Grzech, A., Pieczynska, A., Katarzyniak, R.P.: Agent-Based Approach for Distributed Intrusion Detection System Design. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 224–231. Springer, Heidelberg (2006)
13. Vybiral, D., Augustynek, M., Penhaker, M.: Devices for Position Detection. Journal of Vibroengineering 13(3), 531–535 (2011)

14. Mikulecky, P.: Remarks on Ubiquitous Intelligent Supportive Spaces. In: 15th American Conference on Applied Mathematics/International Conference on Computational and Information Science, Univ. Houston, Houston, TX, pp. 523–528 (2009)
15. Brad, R.: Satellite Image Enhancement by Controlled Statistical Differentiation. In: Innovations and Advances Techniques in Systems, Computing Sciences and Software Engineering, International Conference on Systems, Computing Science and Software Engineering, Electr. Network, December 03-12, pp. 32–36 (2007)
16. Choroś, K.: Further Tests with Click, Block, and Heat Maps Applied to Website Evaluations. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS(LNAI), vol. 6923, pp. 415–424. Springer, Heidelberg (2011)
17. Beben, A., Burakowski, W.: On improving CBR service playback buffer mechanism in WATM networks. European Transactions on Telecommunications 12(4), 311–320 (2001), doi:10.1002/ett.4460120408
18. Chai, W.K., Wang, N., Psaras, I., Pavlou, G., Wang, C.J., de Blas, G.G., Salguero, F.J.R., Liang, L., Spirou, S., Beben, A., Hadjioannou, E.: CURLING: Content-Ubiquitous Resolution and Delivery Infrastructure for Next-Generation Services. IEEE Communications Magazine 49(3), 112–120 (2011), doi:10.1109/MCOM.2011.5723808

Chapter 17
PowerChalk: An Adaptive e-Learning Application

Dan-El Neil Vila Rosado, Margarita Esponda-Argüero, and Raúl Rojas

Abstract. This chapter presents a new interactive e-learning application called PowerChalk. PowerChalk has arisen as the result from the analysis of the evolution of Information Systems Design Theory for E-Learning; it was designed to resolve an important limitation of current design methods and e-learning systems: adaptability. Modular programming is the design technique used in PowerChalk to improve human computer interaction with the management of different types of data in order to have positive effect on both learning score and learning satisfaction. PowerChalk works like a Transaction Processing System in order to support collaboration, communication, creativity and learning through a collection of organized modules. The characteristics of PowerChalk facilitate developing of competencies for using multimedia technologies in any learning session, taking into account the teacher and student perspective. The goal of PowerChalk is to provide a robust, reliable, usable and sustainable multimedia technology for collaborative learning.

17.1 Introduction

The importance of information, multimedia, communication and e-learning technologies in order of promote open, distance, flexible learning satisfaction is obvious; but an important problem not well-documented on e-learning is how and with what resources we need to develop an e-learning application to ensure usability and accessibility to the users.

Learning theorists justify that to reach an objective, acquire a skill or learn something, the learner must be actively involved through practice to cognitively

Dan-El Neil Vila Rosado · Margarita Esponda-Argüero · Raúl Rojas
Freie Universität Berlin, Department of Mathematics and Computer Science
Takustrasse 9, 14195 Berlin
e-mail: {vila80,esponda}@inf.fu-berlin.de,
raul.rojas@fu-berlin.de

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 179–188.
springerlink.com

incorporate it into long term memory [1]. So, the most important characteristic of an e-learning application is the interactivity. Referring to interactive media tools, the adjectives "superior" or "best" depend on the specific context but the interactive whiteboard (IWB) report potential pedagogical benefits and less drawbacks for teachers and students [2]. In general, a formative application should; be interactive and provide feedback, have specific goals, motivate, communicate a continuous sensation of challenge, provide suitable tools, and finally avoid distractions and factors of nuisance interrupting the learning stream [3]. On the other hand a set of features specific for e-learning systems interfaces are: they have to provide a comprehensive idea of content organization and of systems functionalities, simple and efficient navigation, advanced personalization of contents and clear exit.

To accomplish all these characteristics is a difficult task, but through time many e-learning applications have been developed to satisfy some specific goals, however evidence suggests that the solutions are limiting the incentive to innovate and they are restricting the ability to integrate with other systems [4]. Herewith there is a need for theory to support the design and implementation of these e-learning systems. In this approach we analyzed the requirements of e-learning and the state of art in Information Systems Design Theory (ISDT) to propose Modular Programming like a software design technique to solve the most important problem for programmers and users of actual e-learning applications: adaptability.

To support this design-science research, this work has been instantiated in an e-learning application that has been used by staff, students and teachers (PowerChalk).

This chapter is structured as follows. We review ISDT information and the related work in section 17.2. Then, in section 17.3 we describe the PowerChalk system and in section 17.4 the different modules of the system. Finally in section 17.5 the implications of our findings and further research are suggested.

17.2 ISDT and Related Work

Actually, the most efficient tool to use and create high richness multimedia materials is the electronic chalkboard.

Among electronic chalkboard applications we find: educational tools, intelligent work-spaces, group decision making tools, graphical visualizations tools, etc. Currently there are few electronic systems and projects that offer a combination of collaboration platforms, interactive chalkboards and displays that enhance any discussion session. In the state of the art we find the followings projects:

- NotePals. Ink-based, lightweight note sharing. UCLA-Xerox [5].
- E-Cognocracy. Democratic system conceived for extract and diffuse the knowledge derived from a group of people.University of Zaragoza [6].
- K-Sketch. Interface design for creating informal animations from sketches. University of Washigton - University of California [7].
- PADDs. Digital documents that can be manipulated either on a computer screen or on paper. University of Maryland [8].
- SMART systems. Company of electronic whiteboards [9].

- E-Chalk. Electronic chalkboard developed by FU-Berlin [10].
- Cabri software. Interactive media tool to create content faster to accompany text books or provide activities as resources in 2 or 3 dimensions [11].

The above systems are specializing in a very specific task but focusing on availability and usability. They have different limitations depending on stakeholders, among which we mention: costs, hardware or software limitations, their compatibility with only certain types of data and difficulty in updating to different kinds of hardware. In conclusion, they have difficulty in adapting to different kind of stakeholders or circumstances.

The ability to integrate with other systems and evolve are not well satisfied, but Information Systems Design Theory (ISTDT) provides a sight on structures and processes to effectively implement technology for learning activities and improve actual e-learning systems.

David Jones defines three generations of ISDT formulation [12].

- First generation (1996-now). Typified by the generation of requirements, use of templates, software wrappers and commercial off-the-shelf products. Advantages: ability to adapt to change, platform independence. Disadvantages: High level of technical skills required for the users and developers. Results: A growing number of stakeholders feeling limited by the approach, therefore this kind of platforms are abandoned for the users.
- Second generation (1999-now). Delineated by increase use of the system by modifying the development and support processes with insights from diffusion theory, design patterns and pattern mining. Advantages: Instantiations more acceptable to users leading to greater adoption. Disadvantages: The systems have a pro-innovation bias that, amongst other effects, can decrease flexibility and increase difficulties to implement changes. Results: The systems are not able to evolve as quickly as hardware technology, requirements or needs.
- Third generation (2000-now). Differentiated by increase the agility of the system to change by encompassing features from emergent and agile development methodologies and use of OO and patterns. Advantages: Simple design, coding standards and collective code ownership. Disadvantages: Need to coordinate an efficient and disciplined development team. Results: Significant increase in systems use, a notorious decrease of developer's teams and therefore less capability to evolve.

Actual design methods rely on some of the ISDT generations, but in general rely on development being performed by a development team whereas the original template system provided methods by which end users can develop new templates. This characteristic could improve the system's ability to support faster response to change. In this sense PowerChalk is the platform that solve many of the limitations of current systems in order to support diversity, easy development, adaptability and improve human-computer interaction for the management of different types of information.

17.3 PowerChalk Structure

PowerChalk is a collaborative e-learning system for a new kind of electronic chalk, where we can combine the advantages of the traditional chalkboard with the functionality of multimedia, electronic devices and modern distance education tools. PowerChalk will transform any working session into a visual and reliable communication tool.

Goals of PowerChalk:

- To make the system robust, reliable, usable and sustainable with an efficient software structure.
- Provide the platform with a set of tools to build new modules that allow the end-users to analyze complex miscellaneous information quickly, insert relevant notes, access maps and integrate specialized simulation modules with ease (Rich Client Platform).
- Provide the PowerChalk with a communication module via internet to share I formation and have real-time collaborative sessions.
- Easy to adapt to different kind of hardware.
- Easy to update.

With this, we can reach a high-performance system for teaching and learning. For this purpose, PowerChalk was built through a distributed development model based on modularization.

A modular application like PowerChalk is composed of smaller, separated chunks of code that are well isolated. Those chunks of software can then be developed in separated teams with their own life cycle, their own schedule. The results can then be assembled together by a separate entity [13]. This modular architecture has the followings advantages:

- It simplifies the creation of new features. These features can be Macros or useful objects for a multimedia lesson-planning session.
- It makes it easy for users to add and remove features. With this characteristic the user can modify the different tools being used. For example: the electronic ink, the pdf viewer, the image reader and more between the availables modules for PowerChalk.
- It makes it easy to update existing features.

With these benefits PowerChalk becomes a modern, flexible, technologyfriendly approach to e-learning and teaching. The architecture of PowerChalk is showed in Figure 17.1.

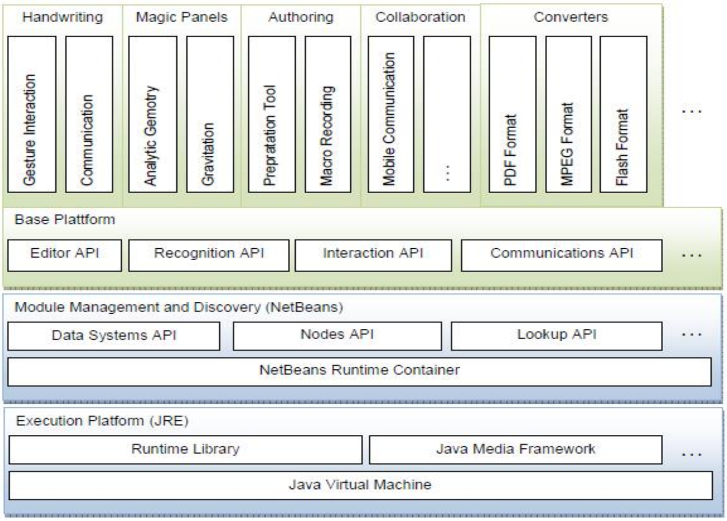


Fig. 17.1 PowerChalk structure

We have a platform and architecture for distributed development with a modularized architecture in Java NetBeans platform. We solved the design problems of another analyzed chalkboard applications through design patterns like navigation, composition, semantic zooming, lookup, etc. [14]. The software structure allows us to give more functionality such as affine transformations over the strokes or images, zooming in all canvas section, layering, distributed development, etc.

17.4 PowerChalk Modules

As applications become more complicated and we need high sustainable software, they are more frequently assembled from modules that were developed independently. In the PowerChalk system, the modules work together to improve our experience as learner or teacher.

17.4.1 Main Editor Module

The base of an electronic board is an electronic ink. We have developed a prototype of an Editor in Java to add electronic ink components, others objects (images, keyboard input, etc.) and its edit functionalities (Figure 17.2). This editor is the base of a handwriting recognition system. Also, the editor includes pen-based applications to process the different kinds of objects through annotation, correction, condensation, organization, zoom abilities, print options (normal and pdf converter) and building of slides.

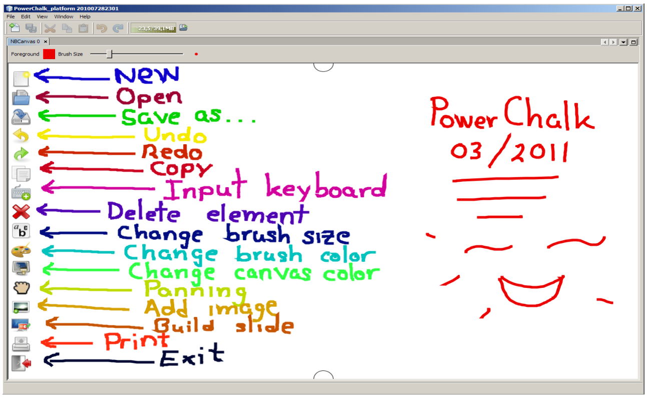


Fig. 17.2 PowerChalk main editor

Teaching a lesson with interactive media involves underlining text, highlighting, commenting or adding other objects on the fly. This combination of showing the information with critical thinking is useful for interactive media learning or active reading. For this purpose, we have an annotation layer mechanism. PowerChalk has a hierarchical mechanism for supporting identification and processing of multiple overlapping layers of annotations for data (images, text, strokes, etc.). It should be noted, that every object in the PowerChalk canvas has its own timestamp and the complete session can be stored. Therefore, archived sessions can be played on-demand as conventional videos, fast-forwarding or rewinding the file. Also, PowerChalk can generate a printable version of the board content via printer or pdf file.

17.4.2 Pen and Digitizer Tablets Module

PowerChalk has a library for accessing pen/digitizer tablets and pointing devices using Java. Its key features are Event/Listener architecture and the fact that device access is implemented through providers in different operating systems and hardware devices (Linux, Windows, Wacom tablets, Hanvon tablets, etc.). Editor tools that interact with the pen tablet have also been developed; an example of this is to open a color chooser through a click with pressure or a selection tool (Figure 17.3). PowerChalk has been used in PC computers and tablet-PC systems for classes in FU-Berlin by teachers and students to test the functionality and practicality.

17.4.3 Multi-screen Manager Module

Rapid adoption of digital devices leads to use several screens to perform the same activities. Every kind of screen has unique benefits; therefore, together they enhance the user experience. In general, the user wants relevant, consistent and connected information across their screens. This module allows the use of PowerChalk in several end-users hardware configurations giving accessibility and increasing productivity in any collaborative learning session (Figure 17.3).

The principal experiments with the Power Chalk Editor are in a Data-wall with four screens that was built in a classroom for testing. We are gathering information from teachers and students to improve the PowerChalk Editor as a part of the usability test.

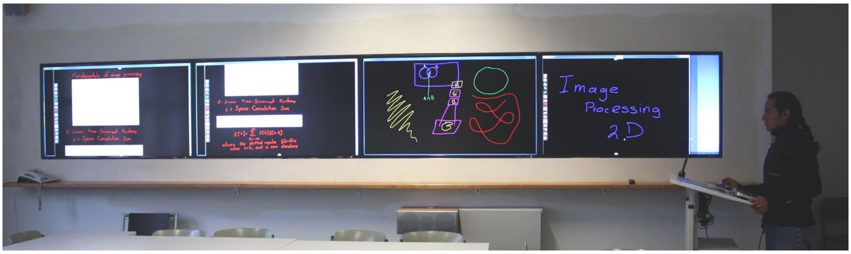


Fig. 17.3 PowerChalk in a multi-screen system with tablet

17.4.4 Handwriting Recognition Module

Adding handwriting recognition to the system makes the electronic board a great tool for a learning session because obtained signal is converted into letter codes which are usable within computer and text-processing applications (Recognition of mathematical expressions, diagrams, etc.) or user interface prototypes. With the handwriting module, we are able to:

- Build bridge modules between PowerChalk and MATLAB,MATHEMATICA, etc. With this functionality, the user just needs to type user interface prototypes for calculations in Mathematica´s language as well as commands or actions for graphical output and the result can be displayed on the PowerChalk canvas (Collaboration modules for applications).
- Recognize user interface prototypes. Pen, finger, and wand gestures are increasingly relevant to many new user interfaces for tablet PCs or PDAs. Designing and implementing gesture recognition was our goal to give PowerChalk a highly human computer interaction tool.

The problem of gesture recognition is divided in 2 cases: one-stroke recognition and multi-stroke recognition. In the one-stroke recognition we have analyzed and implemented methods for use in rapid prototyping [16]. The recognition results for the best configuration for one-stroke recognition reached 1.15% recognition errors (S.D. = 3.69) with 12 training examples. Equivalently recognition rate was 98.85%. For multi-stroke recognition the method worked with 9 training examples and reached a recognition rate of 91.02% and equivalently reached 8.98% recognition error (S.D. = 6.62) [17].

17.4.5 Communication Module

PowerChalk is a system that helps users communicate and collaborate for a certain project. In a PowerChalk session the conversation and information on the chalkboard are equally important. We want the user to be able to communicate and work together with richly formatted text, photos, videos, maps and more.

The structure of communication is through 3 channels: audio, video and PowerChalk canvas data. The system was tested transmitting over RTP and Java Media Framework. The system was improved with an external synchronization.

With this module in PowerChalk we can capture, playback, stream, and transcode multiple media formats providing a powerful toolkit to develop scalable, crossplatform communication technology.

The hierarchy classes of the objects or data (based on Piccolo 2D) allow us to send any object through internet and to show the same information in any other PowerChalk canvas [18]. This technology will allow us to transmit the PowerChalk canvas data with high efficiency and synchronization with the other channels. Otherwise, designing a web framework, web services and an update center for the PowerChalk platform will allow the users to achieve distributed software development for increasing the applications of PowerChalk platform and also, review, edit and build shared sets of recorded sessions. A recorded session could be a lecture, a discussion session over a set of data for group decision-making, an animation, or a set of data processed for gathering information. The communication module for developers is responsible for supporting these activities.

PowerChalk works with a proper mark-up/logic separation, a POJO data model, and a refreshing lack of XML. With this kind of application, the end-user is able to publish any work session to a general public, for example: the classroom notes or homeworks.

Web services are software systems that are externally available over a network. You can use them to integrate computer applications that are written in different languages and run on different platforms. Web services are independent from language and platform because vendors have agreed on common web service standards. PowerChalk works with RESTful Web Services and SOAP-based Web services.

Running the Update center will check if there are new modules or new versions of already installed modules available. From new or updated modules, the user could select, download, and install them. We developed an update center module for the PowerChalk structure over the HTTP protocol.

17.4.6 Collaboration Modules

A collaboration module for applications is a bridge module between PowerChalk and software like MATLAB, OCTAVE, MATHEMATICA, Gnuplot, etc. With this kind of modules we can send instructions for plotting functions, expression evaluation, solving equations, running a script, etc. and see the output on the PowerChalk canvas. We have designed an API for this kind of modules. The call to Gnuplot, or another application is made trough a keyboard input or handwriting information in the PowerChalk canvas.

17.4.7 Magic Panels

A magic panel module has an area on the PowerChalk canvas, where the users can embed their own objects, tools or applications. An example is the animation module. An animation is one of the best ways to express moving visual images because it can represent dynamic concepts and it can make information more attractive and engaging however making animation is out of reach for the common user. Most animation tools are complex and time consuming to learn and use.

To endow PowerChalk with an informal animation tool we merge the structured 2D graphics framework Piccolo2D [19] with an efficient lookup system. The design of the object structure with the lookup system in PowerChalk allow us to implement commands to animate objects over the PowerChalk-canvas, for example, we can animate affine transformations over an object and composite a set of elements. As a result, PowerChalk has an API for animations and the platform for development. The user will be able to build a magic panel.

17.4.8 Macros Module

We use the term Macro to make available to the user, a sequence of quick notes (notes written in a prior time) or information (images, text, videos or pdf files) to use in a class session. The user interface of PowerChalk allows having the information for a class, available in an organized and efficient window system.

It should be noted that we are able to add to the macro notes or a page of notes written in a normal paper. This module can export a scanned file of the notes, binarize the image, make the segmentation of different annotations (if the annotations are split for a horizontal line) and make it available to the macro.

17.5 Conclusions and Future Work

PowerChalk preserves the pedagogical benefits of the traditional chalkboard and provides the possibility to show not only results or isolated ideas but the train of thoughts. Communication was established with some universities, companies, researchers and students for testing the capability to develop and perform teaching sessions, and also the usability and functionality of PowerChalk. The consensus is that PowerChalk it is a robust, reliable, usable and sustainable score learning platform and also a friendly tool to review and share lecture notes. To support in a measureable way the advantages of PowerChalk, a complete usability test is in progress. Also, we are increasing the efficiency of every module.

The modular structure of PowerChalk allows us to quickly amend any detail of the system since PowerChalk has the distributed development characteristic. Because of this advantage new modules are being developed. Among the future modules we can find: Voice recognizer module, image processing module and a java compile module, etc. The idea is to transform the PowerChalk system (e-learning system) into a complete platform to share and process general information.

References

1. Dick, W., Carey, L., Carey, J.O.: The Systematic Design of Instruction, 6th edn., pp. 1–12. Allyn & Bacon (2005) ISBN 0205412742
2. Twiner, A., Coffin, C., Littleton, K., Whitelock, D.: Multimodality, orchestration and participation in the context of classroom use of the interactive whiteboard: a discussion. Technology, Pedagogy and Education 19(2), 211–223 (2010)
3. Norman, D.: Things that make us smart: defending human attributes in the age of the machine. Perseus Publishing, Cambridge (1993)
4. Anonymous: Mixed skies ahead: What happened to e-learning and why. The Learning Alliance for Higher Education. Change, March-April (2004)
5. Davis, R.C., et al.: In: Proceedings of the SIGCHI Conference on Human Factors Incomputing Systems (1999)
6. Lanuza, A.T., et al.: REDALYC: Magazine Computación y Sistemas 12(2), 183–191 (2008)
7. Davis, R.C., Colwell, B., et al.: K-sketch: A "kinetic" Sketch Pad for Novice Animators (2008)
8. Guimbretiere, F.: Computer Human Interactions Letter, vol. 5(2), pp. 51–60. ACM (2003)
9. `http://smarttech.com/`
10. Jantz, K., Friedland, G., Rojas, R.: Ubiquitous Pointing and Drawing. International Journal on Emerging Technologies for Learning, iJET (2007) ISSN: 1863-0383
11. `http://www.cabri.com`
12. Jones, D., Gregor, S.: The formulation of an Information Systems Design Theory for E-Learning. In: Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (2006)
13. Boudreau, T., Tulach, J., Wielenga, G.: Rich Client Programming, Plugging into the NetBeans Platform. Prentice Hall (2006) ISBN: 0-13-235480
14. Gamma, et al.: Design patterns: Elements of reusable object oriented software. Addison-Wesley professional computer series (1994)
15. Wobbrock, J.O., Wilson, A.D., Yang, L.: Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In: Proceedings of the ACM Symposyum on User Interface Software and Technology (2007)
16. Anthoy, L., Wobbrock, J.O.: A lightweight multistroke recognizer for user interface prototypes. In: Proceedings of Graphics Interface 2010 (2010)
17. Bederson, B., Grosjean, J., Meyer, J.: Toolkit Design for Interactive StructuredGraphics. IEEE Transactions on Software Engineering (2004)
18. `http://www.piccolo2d.org`

Chapter 18
Identifying Fraudulent Shipments through Matching Shipping Documents with Phone Records

Czesław Jędrzejek, Maciej Nowak, and Maciej Falkowski

Abstract. A crucial element of a VAT-carousel crime scheme is false documentation for transporting many goods (fuel, steel rods); claims that trucks cross the Polish border to and from neighboring countries (operation entailing zero-rate VAT tax), while, in reality, they never leave Poland. It is possible and practical to use analytic technology to prove that the routes determined by the locations of radio towers of drivers' calls are incompatible with alleged transport documents (CMRs') and invoice based information. This is based on a rule based expert system that covers semantic matching, spatiotemporal relation operators, and comparison of GIS data. The method has been successfully applied to real-life investigation (the GARO case).

18.1 Introduction

Several fraudulent schemes use the right to deduct input VAT for fictitious transactions. In this chapter we analyze the following crime scheme (the Polish version of VAT-carousel crime) represented in Fig. 18.1.

Czesław Jędrzejek · Maciej Nowak · Maciej Falkowski
Institute of Control and Information Engineering
Poznan University of Technology, Poland
e-mail: czeslaw.jedrzejek@put.poznan.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 189–198.
springerlink.com

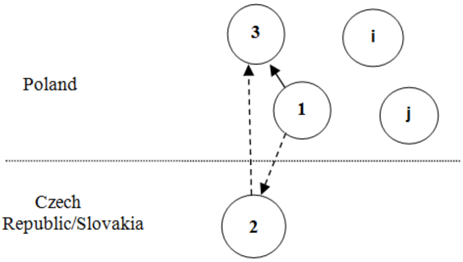


Fig. 18.1 Companies involved in a crime scheme

There are two possible areas of investigation.

1. Financial flows and tax information.

This is difficult although the most important. Since many straw companies appear in a scheme, Tax Offices and police waste a lot of time and effort investigating of all companies; this has minimal benefit, compared to going after masterminds and recovering money from fraudulent VAT-tax reclaims. This is a topic of our parallel work [1].

2. Data correlation between invoices and transnational shipment documents, (**C**ontrat de transport international de **M**archandise par **R**oute, CMR documents are used in this chapter) and driver's phone records. Cell phone data map the positions of trucks along the routes they have taken (as recorded on radio towers, called BTS, Base Transceiver Stations). These phone records based routes can be matched against with these stated in CMR documents (complemented by invoices).

In this work we concentrate on the second aspect of the issue. The falsification of routes is a prerequisite to VAT carousel crime, because VAT tax on transnational sale has zero rate. In the particular type of VAT-carousel crime, the analysed shipment pattern is the following:

- Company 1 from Poland issues an invoice and international transportation document, CMR, related to a (fictitious) sale of goods t (here steel rods) to a company 2 in the Czech Republic/Slovakia.
- Company 2 then sells the goods t to Poland and issues an invoice and international transportation document, CMR, related to a (fictitious) sale of goods t (the same steel rods) to a company 3 in Poland.
- In reality, the goods t do not leave Poland. They are sold by company 1 to company 3 through a series of intermediaries i.

The method has been successfully applied to real-life investigation concerning fraudulent process of transaction between Polish and Czech Republic companies. This investigation is called the GARO (anonimized acronym) case.

The chapter is organized as follows. Section 2 presents a model of data correlation between invoices, CMR documents and drivers' phone records. Section 3 describes a route verification algorithm. Section 4 shows results for the GARO case. Conclusions and future work are presented in Section 5.

18.2 A Model of Data Correlation between Invoices, CMR Documents and Driver's Phone Records

Uncovering criminal behaviour relies on comparing data related to fictitious and real routes taken by shipping companies and drivers. Depending on the distance between origin and destination locations of goods, as well as data on dates of transportation and frequency of phone calls, a meaningful percentage of routes can be labelled as fictitious. At this moment at least 20 cases are being investigated, the GARO case being most representative of the scheme.

The most relevant shipment documentation (CMR records) data are pairs:

{X_1 – route origin in Poland, T_1 – date of transport from Poland }

{X_2 – alleged destination in the Czech Republic/Slovakia, T_2 – goods receiving date in the Czech Republic/Slovakia}

{X_3 – actual destination in Poland, T_3 – goods receiving date in Poland}

From shipment documentation (CMR records), the alleged routes in the GARO case were the following:

X_1 ={Ostrowiec Swietokrzyski, Zawiercie, Zawada}; Zawada is the truck base of the EI company,

X_2={Bystrice, Cadca},

X_3={ X_i , Zawada}, where X_i are locations of end clients.

It many cases parts of the alleged routes $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_3$ are common, which makes analysis much more difficult.

As for the set $\{T_1, T_2, T_3\}$ in most cases $T_1 = T_2$ and T_3 is missing. Missing T_1 does not prevent reasoning, the deliberate omission of data makes it impossible to establish truth. In some cases the missing data could be obtained by inspecting additional documents, but no attempt has been made to do so.

Phone calls are represented in the form of phone records $PR\{PN_k, T_s [l_1,t_{1;}\ l_p,t_{p;}\ l_N,t_N]\}$, where:

PN_k = phone number k belonging to a driver D_k,

T_s = day of transport,

l_p,t_p = pair - a location and a time of start of a phone connection on a given day T_s.

In further analysis the following simplifications are assumed:

1. There exists only one phone number k used by a driver D_k. Moreover, drivers do not exchange terminals or SIM cards between themselves.

2. The same driver drives a truck along fictitious routes $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_3$ as well as the real route $X_1 \rightarrow X_3$. In the newest cases sometimes one driver rides an empty truck to the Republic/Slovakia and picks up a load left at the warehouse by another driver.

   These two assumptions occurred in the GARO case.
3. The algorithm in the present form does not take into consideration date T_1. Categorization is based on knowledge of T_2 and T_3. If a driver could have been in a warehouse before 15 o'clock and T_3 is not known, we assume that $T_3 = T_2$; however, the analysis becomes less conclusive.

   It is also known that in some newer cases legs of $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_3$ are, according to documents, completed by different drivers.

Initially, the police collect CMR (and invoice documents) in a paper form for a given period, and transform them to an Excel files containing relevant data; this plus electronic phone records of drivers constitute a database for a given case.

Analysis is performed for the whole time period covered by CMRs and invoices, processing documentation for successive dates from a list of CMR documents. For a given date T_2 and a given driver, a matching phone record set is selected for the same day. The phone record set is sorted with increasing time.

We used GoogleMaps for a route and travel speed determination. Because of uncertainty regarding the speed of travel and the positions of driver within the range of base stations (BTS) towers to compensate for this we assumed that a transport was serviced if it reached Bystrice before 15 o'clock.

Inspecting the phone record location statistics for 3 drivers showed that none of 2366 analysed connections was routed by an operator from the Czech Republic/Slovakia

18.3 A Route Verification Algorithm

We have developed 3 criteria relevant for determining whether a shipment to the Czech Republic/Slovakia could occur as stated in CMR documents.

1. Whether the driver could have reached the destination point in the Czech Republic/Slovakia by 15 o'clock on day T_2
2. Whether the driver could have reached the destination point in the Czech Republic/Slovakia on day T_2
3. Whether the driver could have reached the goods receiving location in Poland on day T_3, if known.

The negative answer to any of these questions indicates that the owner of the shipping company and drivers that carried out transport fall under sanctions from Art. 271 § 3 of the Polish Penal code (PC) – lying or stating (signing) a falsehood in a document with regard to a circumstance having legal significance. What is more important, this constitutes a prerequisite to a fraudulent VAT reclaim, which may be further investigated by police or Tax Office officers.

The algorithm used in our rule-based expert system is based on semantic matching, spatiotemporal relation operators, and visualization of GIS data.

A very important concept is a span of the time window between two consecutive calls, that is analysed with regard to spatial rules.

Extensive work exists on this issue [3-5]. In previous work various ontologies and rules were proposed, written in formalized frameworks and rule languages such as RuleML and SWRL.

In our case, geographic locations come from CMR documents and phone records. In the spatiotemporal representation used in this work, the most important aspect is time ordering. The essence of our algorithm is calculation of whether a driver could have travelled distance along the route $y_1 \rightarrow X_2 \rightarrow y_2$, (where $y_1=l_p$ and $y_2= l_{p+1}$), passing though the warehouse in the Czech Republic/Slovakia, X_2, in a given time span. Use of spatial ordering is not very effective for the purpose of building evidence, because, in general, we cannot determine whether a point l_p was achieved on the way to X_2 or in the reverse direction. The driver could drive in a senseless zigzag fashion; however sometimes this pattern could have made sense, particularly, if he made two shipments in a given day. Without additional information, we chose not to employ special spatial ordering rules. We introduce the following notation:

RPR - Reference Phone Record, the one having $t_{p,}$ such that $[(t=15:00) - t_p] >0$, and is minimum.
DLPR - Direct Later Phone Record, i.e. having time t_{q+1} with regard to t_q.
DEPR - Direct Earlier Phone Record, i.e. having time t_{q-1} with regard to t_q.
FDNPR - Following Direct Later Phone Record, i.e. having time t_{r+2} with regard to t_r.

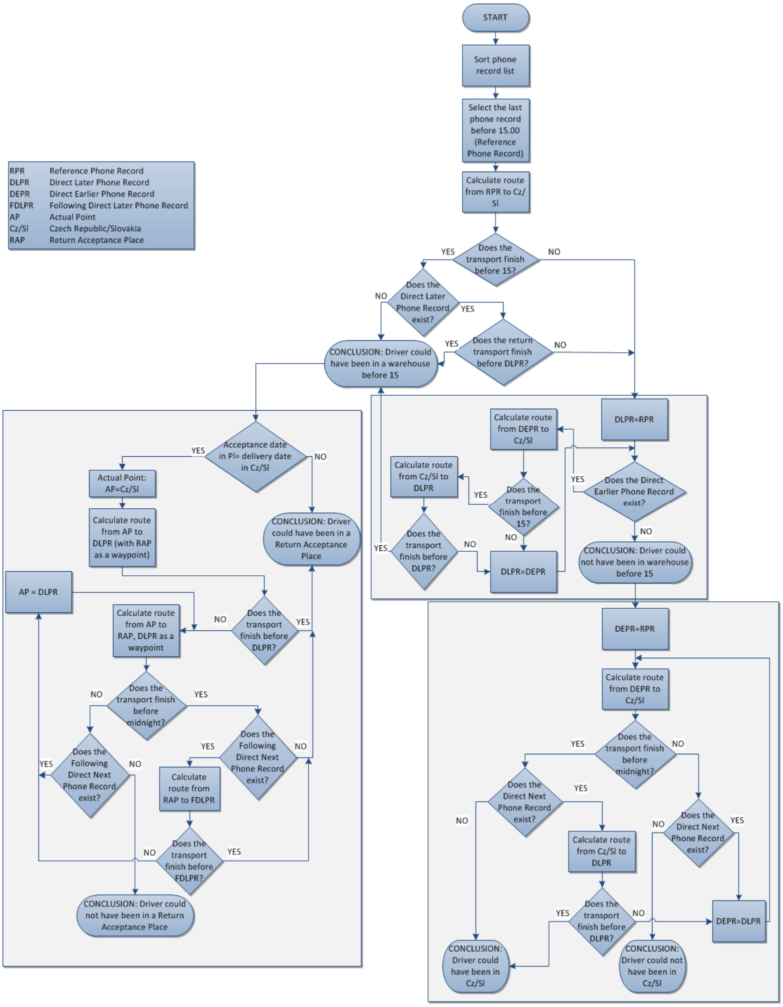
These objects have two attributes: location and time, however, in the flowchart determining outcome of analysis (in Fig. 18.2), we use names of the object instead of their spatial attributes in order to save space in the figure.

AP - Actual Point { X_2 or location l_p from PR{ $l_p,t_{p;}$; $t_p >t(X_2)$}; where $t(X_2)$ is the calculated time of reaching X_2.
Pl – Poland,
Cz/Sl -Czech Republic/Slovakia,
RAP - Return Acceptance Place.

We attempt to get conclusions on three criteria relying on the truthfulness of statements on CMR documents

1. Driver could have been in a warehouse before 15 o'clock on day T_s.
2. Driver could have been in Cz/Sl on day T_s.
3. Driver could have been in goods receiving place in Poland on day T_t.

The decision flowchart is drawn in Fig. 18.2. The shaded blocks correspond to 3 criteria: the left block probes criterion 3. The top right block probes criterion 1 and the bottom right block probes criterion 2.



RPR	Reference Phone Record
DLPR	Direct Later Phone Record
DEPR	Direct Earlier Phone Record
FDLPR	Following Direct Later Phone Record
AP	Actual Point
Cz/Sl	Czech Republic/Slovakia
RAP	Return Acceptance Place

START
Sort phone record list
Select the last phone record before 15.00 (Reference Phone Record)
Calculate route from RPR to Cz/Sl
Does the transport finish before 15?
YES
NO
Does the Direct Later Phone Record exist?
NO
YES
Does the return transport finish before DLPR?
YES
NO
CONCLUSION: Driver could have been in a warehouse before 15
Acceptance date in Pl= delivery date in Cz/Sl
YES
NO
Actual Point: AP=Cz/Sl
Calculate route from AP to DLPR (with RAP as a waypoint)
CONCLUSION: Driver could have been in a Return Acceptance Place
AP = DLPR
Calculate route from AP to RAP, DLPR as a waypoint
Does the transport finish before DLPR?
NO
YES
Does the transport finish before midnight?
NO
YES
Does the Following Direct Next Phone Record exist?
YES
NO
Does the Following Direct Next Phone Record exist?
YES
NO
Calculate route from RAP to FDLPR
Does the transport finish before FDLPR?
NO
YES
CONCLUSION: Driver could not have been in a Return Acceptance Place
DLPR=RPR
Calculate route from DEPR to Cz/Sl
Calculate route from Cz/Sl to DLPR
Does the transport finish before 15?
YES
NO
Does the Direct Earlier Phone Record exist?
YES
NO
Does the transport finish before DLPR?
YES
NO
DLPR=DEPR
CONCLUSION: Driver could not have been in warehouse before 15
DEPR=RPR
Calculate route from DEPR to Cz/Sl
Does the transport finish before midnight?
YES
NO
Does the Direct Next Phone Record exist?
NO
YES
Calculate route from Cz/Sl to DLPR
Does the Direct Next Phone Record exist?
NO
YES
Does the transport finish before DLPR?
YES
NO
DEPR=DLPR
CONCLUSION: Driver could have been in Cz/Sl
CONCLUSION: Driver could not have been in Cz/Sl

Fig. 18.2 The route classification algorithm

The transport classification algorithm checks all the time windows - periods of time between two phone calls of the driver, to determine if the driver could have completed a route to a warehouse in the Czech Republic/Slovakia and back. Since we know a location of each connection (operators provides data on the BTS towers which registered a connection), these points are used as the start and the end of the analysed route.

Time of departure from the place of loading in Poland is not known, and therefore the algorithm does not assume any. In many cases, T(X_3), the date of the return of truck to Poland, is not known; the suspects deliberately omit it.

We then assume that the supposed return shipment to Poland took place on the same day t(X_2). This is a suggestion to possibly complement the existing case database with additional data: from road inspections, or tachometer recorders. Introduction of several possible options for a day of return would significantly increase the number of possible cases to be analyzed (also generated drawings), and increase number of rules. Initial results indicated no significant evidential benefit.

The architecture of the created environment is sketched in Figure 18.3. The input data in the form of Excel tables is entered into Palantir Graph [6]. Palantir represents data as objects, in this case in the form of a phone call, and route origin and destination events. Important information is selected from data initially visualized in this way by the set of object filters. Filtered data is geolocated, and placed on Palantir Map application. Route Classifier separates all objects according to their dates, and calculates whether the documented transport routes are possible or not. As the result the analyst gets a set of pictures with drawn routes and phone calls placed on their respective BTS towers, each representing one analysed day.

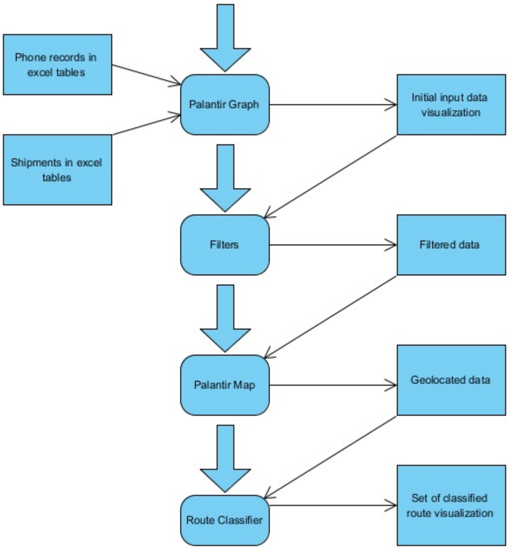

Phone records in excel tables
Shipments in excel tables
Palantir Graph
Initial input data visualization
Filters
Filtered data
Palantir Map
Geolocated data
Route Classifier
Set of classified route visualization

Fig. 18.3 Architecture and flow of data in the analysing environment

The input data in form of excel tables consists of two types of information:

- phone records - list of phone calls made or received by the given driver. Each phone record has the BTS tower id, which indicates where it was recorded.
- shipment documents - information about dates, origin, destination and type of transport, as well as CMR containing driver's data.

18.4 Results for the GARO Case

We have applied our method to the real data related to the GARO case; phone records pertain to the period of 11 months between July 2009 and the end of April

2010. In Table 18.2 we present results related to three drivers K.K., R.K. and G.K. working for the EI company. Statement “assuming that the date is the same” means $T_3=T_2$.

Table 18.1 Categorization of case showing levels of data falsification in CMR documents. The last 3 columns present a number of category assignment for three drivers

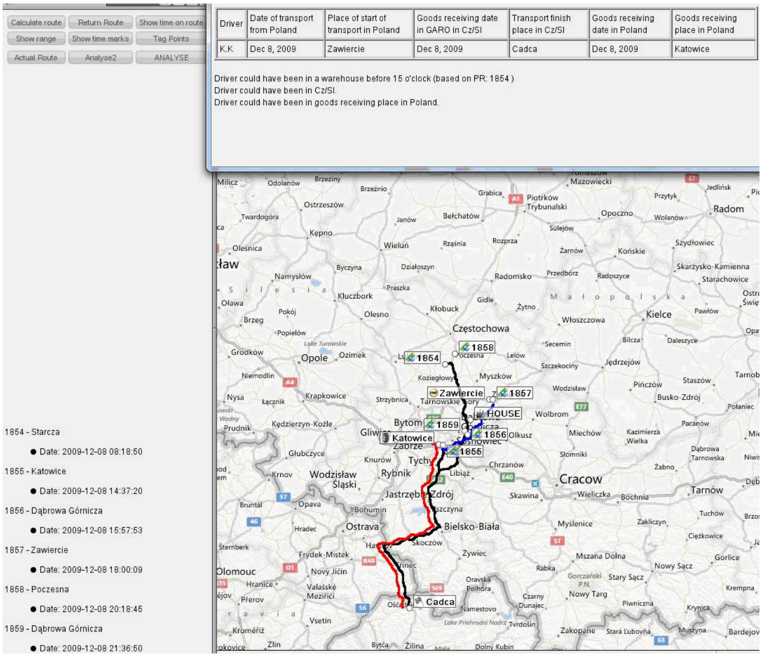
Category	K.K.	R.K.	G.K.
1. Driver could have been in a warehouse before 15 o'clock. Driver could have been in Cz/Sl. Driver could have been in goods receiving place in Poland.	1	0	3
2. Driver could have been in a warehouse before 15 o'clock. Driver could have been in Cz/Sl. Driver could not have been in goods receiving place in Poland.	4	3	3
3. Driver could not have been in a warehouse before 15 o'clock. Driver could have been in Cz/Sl. Driver could not have been in goods receiving place in Poland	23	37	20
4. (including cases assuming that the date is the same).			
5. Driver could not have been in a warehouse before 15 o'clock. Driver could not have been in Cz/Sl. Driver could not have been in goods receiving place in Poland.	35	21	41
6. Driver could have been in a warehouse before 15 o'clock. Driver could have been in Cz/Sl. Driver could have been in goods receiving place in Poland (assuming that the date is the same).	3	4	10
7. Driver could have been in a warehouse before 15 o'clock. Driver could have been in Cz/Sl. Driver could not have been in goods receiving place in Poland (assuming that the date is the same).	0	0	0
Missing data prevents categorization	1	0	0
TOTAL	66	65	77

As we see only in a fraction of cases: in categories number 1 and 5 (shaded in green), the drivers could have taken routes as documented. The reverse happens for categories 2-4.

It is important to present routes on the relevant maps. We colour-code the routes $X_1 \rightarrow X_2$ (black), $X_2 \rightarrow X_3$ (red) and the real route $X_1 \rightarrow X_3$ (blue). Initially, all routes, including the route from the place of start of transport in Poland X_1 to the place of transport finish in the Czech Republic or Slovakia X_2, are drawn based on the input documents (CMR from shipment companies). It represents a documentary state of the investigation, which, in most cases, is fictitious.

Next, we verify the documentary $X_1 \rightarrow X_2$ route against the route that is drawn using BTS towers’ location from phone records as waypoints in order to show the supposed route of the shipment. The segments between locations of consecutive BTS towers’ locations are draws using GoogleMaps. The composite route is tested against the set of rules, in order to reason whether it was possible or no to be in the place of transport finish.

When the algorithm finds a time window, in which the transport could have taken place, the route is redrawn, going from $y_1 \rightarrow X_2 \rightarrow y_2$, where y_1 and y_2, are the locations of BTS' towers (treated as as waypoints) of the phone records between which the time window on travelling to the Czech Republic/Slovakia and back exists. The meaning of the black route ceases to be $X_1 \rightarrow X_2$. We were able to write scripts that render the whole procedure in the automatic fashion.



Driver	Date of transport from Poland	Place of start of transport in Poland	Goods receiving date in GARO in Cz/SI	Transport finish place in Cz/SI	Goods receiving date in Poland	Goods receiving place in Poland
K.K	Dec 8, 2009	Zawiercie	Dec 8, 2009	Cadca	Dec 8, 2009	Katowice

Driver could have been in a warehouse before 15 o'clock (based on PR: 1854)
Driver could have been in Cz/SI.
Driver could have been in goods receiving place in Poland.

1854 - Starcza
- Date: 2009-12-08 08:18:50

1855 - Katowice
- Date: 2009-12-08 14:37:20

1856 - Dąbrowa Górnicza
- Date: 2009-12-08 15:57:53

1857 - Zawiercie
- Date: 2009-12-08 18:00:09

1858 - Poczesna
- Date: 2009-12-08 20:18:45

1859 - Dąbrowa Górnicza
- Date: 2009-12-08 21:36:50

Fig. 18.4 Documentary $X_2. X_3$ (red) and the real $X_2 . X_3$ (blue) route for K.K. driver for the transport on December the 8th 2009. The possible $X_1 . X_2$ (black) route could occur between locations of BTS, 1854 and 1855

18.5 Conclusions and Future Work

In this chapter we presented a practical scheme of proving fictitiousness of fabricated shipping documents which is a prerequisite to building evidence against criminals that defraud Government money by cashing the false VAT tax reclaims.

The analysis of shipment documents performed in this work could provide a filter that would prioritize investigations. Moreover, this analysis will be a part of a general model of VAT tax related crimes. The crucial element of analysis was availability of phone records of involved drivers. All telecom firms keep data for a period. Currently, in Poland this period is two years. However, a new regulation is prepared that limits this period to half a year. If this happens law enforcement and intelligence agency investigations will be greatly limited in their ability to provide evidence of associations between individuals and a particular location. In the GARO case most of the phone records used in this analysis would not be available if the extension period were half a year.

Starting late 2012 data from vignette viaTOLL system would possibly be available for truck location analysis.

Another issue is insufficient of cooperation between countries. In the GARO case warehouse were allegedly used outside of Poland. To use this fact as evidence in the Polish court involvement of the foreign countries law enforcement institution would be necessary.

The analysis in this work is a basis for logic enhanced link analysis of companies involved in the scheme [1]. The complete analysis on the scale of Poland would require very large procedural changes in Polish institutions and replacing proprietary tools by cheaper substitutes to make the system more cost effective.

Acknowledgments. This work was supported by the PUT DS 4-45-085/11 DS PB grant.

References

1. Nowak, M., Jedrzejek, C., Bak, J., Szwabe, A.: A rule-based expert system for building evidence in VAT-carousel. In: Multimedia and Internet Systems: New Solutions, Wrocław (2012) in print
2. Gao, S., Boley, H., Mioc, D., Anton, F., Yi, X.: Geospatial-Enabled RuleML in a Study on Querying Respiratory Disease Information. In: Governatori, G., Hall, J., Paschke, A. (eds.) RuleML 2009. LNCS, vol. 5858, pp. 272–281. Springer, Heidelberg (2009) http://www.transportsfriend.org/int/docs.html
3. Ahlqvist, O.: On the (Limited) Difference between Feature and Geometric Semantic Similarity Models. In: Claramunt, C., Levashkin, S., Bertolotto, M. (eds.) GeoS 2011. LNCS, vol. 6631, pp. 124–132. Springer, Heidelberg (2011)
4. Keßler, C., Raubal, M., Wosniok, C.: Semantic Rules for Context-Aware Geographical Information Retrieval. In: Barnaghi, P., Moessner, K., Presser, M., Meissner, S. (eds.) EuroSSC 2009. LNCS, vol. 5741, pp. 77–92. Springer, Heidelberg (2009)
5. http://www.palantir.com/

Chapter 19
Open Personal Identity as a Service

Miroslav Behan and Ondrej Krejcar

Abstract. The mobile technologies establish communication environment where mash able solutions are more than convenient. Open personal identity is independent service which is gathering available identity resources and provides unified person identities. The service enables to resolve current mobile device problematic around multiplicities, backup or change management of person identification where multiple devices replication is an option.

19.1 Introduction

Do you remember the situation when you have changed your phone number and you had to tell this change to all of your friends, relatives even workmates? That time is over with the Open Person Identity as a Service. Imagine worldwide Internet service which provides on-line personal information such as mobile numbers, current living address or current friend's cross different social media. There are many advantages of usage of such a kind of service. We would like to introduce some of them in more details.

The modern knowledge society produces much more information than we are able to consume and therefore the utilization or clarity of information is more than convenient. Only those kinds of services which are not complicated or confusing would be accepted by many and the strength of intuitive factors for applications or services behavior will increase in time. That's why social media have such power of influence because they are gathering information from many sources in easy and comprehensive personal way. The problem is when you have more social media then the amount of time spent by scanning or posting into the different sources would not be efficient. The case is about to find an open solution which consolidates all media in one place and basically provide personal social connector as a convenient user-friendly solution with an easy and comprehensive user interface.

Miroslav Behan · Ondrej Krejcar
University of Hradec Kralove, FIM, Department of Information Technologies
Rokitanskeho 62, Hradec Kralove, 500 03, Czech Republic
e-mail: miroslav.behan@uhk.cz, ondrej.krejcar@asjournal.eu

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 199–207.
springerlink.com

19.2 Problem Definitions

The amount of social media networks, multiplicity of personal identity[2] and the inconvenient way of handling the important personal information lead us to think that there some better ways how to make our lives a little bit easier. That's why we start to think about the problem in terms of usability in current available on-line social technologies [8, 9].

We started to ask how to solve our daily life common problems and we summarized them in following questions. What if we have more than one mobile device but each one of them has a different content? Or if we have just one mobile device but we lost it? Could we exchange mobile device platforms without any inconveniences? Do we have to notify everyone when we change our mobile number or even when we do not use it anymore as an identity? When we answered positively to some of those questions, we considered us in correct problem definition [10, 11].

That was just a brief overview of a complex task to solve. In this article we are focusing on personal identity service which is used for virtual personal identification and enables communication between people over modern technologies; nevertheless we consider that kind of service as open and as an independent concept where commercial influences are minimized. At first we describe communication process between two or more sides where communication could be established if there is an existing compatible informational data flow exchange between mobile device clients. To start process at first we need to know the identity of persons with whom we would like to communicate. The identification of personal identity consists of our tacit knowledge where the identity is located in available informational resources and how is the identity knowledge externalized by visualization in comprehensive form. After correct identification of required person the communication process can start.

As current personal identification mainly used in mobile devices we assume a phone contact list where identities are expressed by names, personal pictures or associated phone numbers. That kind of establishment was made by mobile providers over the world. Another personal identity used in mobile device communication that we recognize are the instant messaging systems where identities are commonly defined by user name coded by sequence of characters. We consider these types of identification as obsolete and we propose a new concept in chapter New Design.

Also we define the environment as an on-line with unlimited access to the Internet according to the fact that the increasing mobile device on-line connectivity is arising. We announce the off-line mode of Internet connectivity as temporary state which is identified by status of not connected client and which would be changed by user interaction or predefined settings device behavior to on-line mode and proceeds in delayed tasks. We considered on-line Internet access to mobile device in terms of synchronization of contact list with the Open Person Identity Service (OPIS) over message based client-server where changes are only made by authorized identity owner. In those terms of change management we defined following concept of Front-End and Back-End where:

Front-End – all clients which are accessing over Internet to service by message based communication and perform user's actions corresponding to correct content within associated devices and also can perform data merge operation with current device content.

Back-End – the server provides service based on client-server type of connection and background resource processing which interacts with social media as automated direct resource connector.

Next chapter highlighted the solution concept (see figure 19.1) which can solve our defined problematic by changing establishment and by exploiting today's technologically innovated environment surrounding us with increasing mobile Internet connectivity.

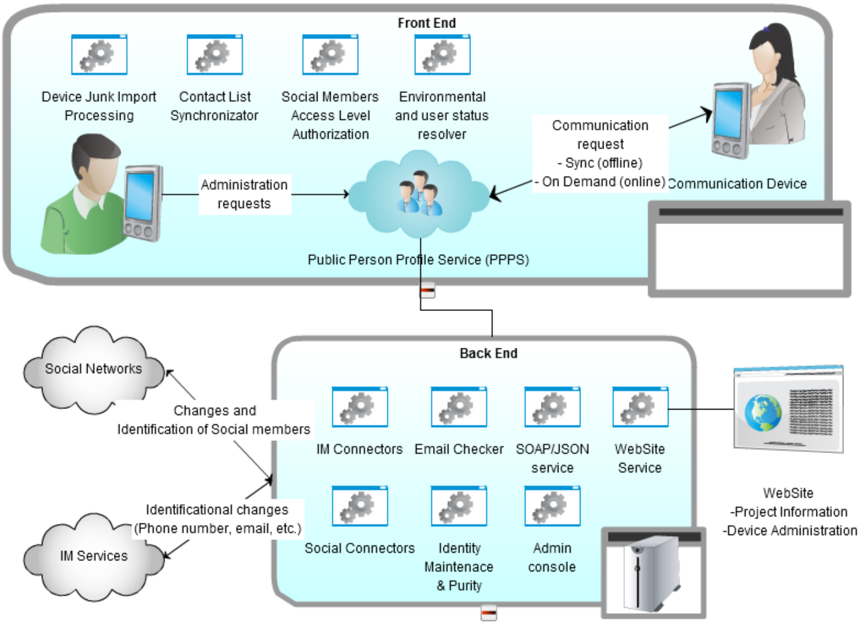

Front End
Device Junk Import Processing
Contact List Synchronizator
Social Members Access Level Authorization
Environmental and user status resolver
Communication request
- Sync (offline)
- On Demand (online)
Communication Device
Administration requests
Public Person Profile Service (PPPS)
Social Networks
Changes and Identification of Social members
Identificational changes (Phone number, email, etc.)
IM Services
Back End
IM Connectors
Email Checker
SOAP/JSON service
WebSite Service
Social Connectors
Identity Maintenace & Purity
Admin console
WebSite
-Project Information
-Device Administration

Fig. 19.1 Scenario of Open Personal Identity as Service

19.3 Related Works

Today's personal identities are stored mostly in mobile device as a contact list saved on a local storage. Synchronization with other devices or with desktop applications is normally made over USB or Bluetooth which is connected directly to personal computer. For instance we just highlight some of software solutions: Apple iTunes, Nokia PC suite, Microsoft Phone Data Manager. These mentioned software solutions have some disadvantages. The installation requires dedicated computer where are all data and management placed. Supported mobile devices are basically only with corresponding platform or manufacturer in terms of single content management or in case of mobile device lost or exchange.

Those disadvantages of current local data management software of mobile devices led us to propose remote data management solution so a part in this article is covering a solution for personal identities service based on a contact list embedded in mobile devices, which could be manageable from device itself or from web interface from Internet [4, 5].

The reason why we considered such solution is a usability of mobile devices due to its limitations in editing the contact where small screen and lower level maturity of a user's input interface is provided in comparison with common desktop. The other reason is a possibility of data replication to other different types of mobile devices. In short it is to create an independent platform for mobile phone users who have more than one device. It is also useful for an easy recovery of a contact list data in case the device is lost or broken.

The new solution considers security issues and authorization of publishable personal information. The main reasons why such a service may be not acceptable from user’s point of view are data privacy issues where users will not like to share data of their contacts. That issue we solve in terms of use and encryption system policy where no one could decrypt personal data without a password.

We announce well-known OpenID service as different type of web service [6, 7] which basically provides uniform access to multiple web sites or application which implements OpenID access as a 3rd side authentication process. The principals are different in basic scenarios where for example in case of OpenID the user visits a web application and is able to log in without registration or native login process but instead of that the user will be only authorized by OpenID with the same credentials when service is implemented and provided. The principals about OPIS are described as following use case scenario. The user have only one place for real identity attributes and these information are in case of change automatically redistributed to connected systems or they are provided as a service like on-line requests by gathered data from social connectors where last update event of specific identity detail is provided.

19.4 Solution Design

As was mentioned in chapter above the developed solution is based on front-end and back-end architecture where as a front-end we assume only devices which are opened to software maintenance and which are configurable such as smart phones, tablets, and computers or even for instance the cars with embedded customizable control unit [3, 11]. The front-end in our perspective is basically any customized client with Internet connection ability, device with contact list accessibility and with background processing possibility. As an extension of front-end in term of user device application also user interface whereas the output we consider graphical (GUI) or voice interface (VUI) and as the input a touch, keyboard or voice recognition user interface [12, 13]. Next part, the back-end could be any server technology which is able to store data of identities and their associations with clients, which have Internet connectivity and providing services on specified ports and also which are able to maintain informational flow between external resource providers such as social networks or instant messaging services and internal website accessibility for remote device administration [14, 15]. That was in short the concept of described solution where we are focusing on types of user actions on client side and then on server side on back-end processing.

For more precise description of front-end we would define common end-user's actions and divide them into two parts as an interoperability types of actions which come first and as an administrative action types which come after. The task that would not necessary starts at first time after client installation is the import of personal identities processing where available resource is embedded in a device contact list, in usage of instant messaging systems or in social networks. All that kind of application would be recognized at first time or upon user's additional task completion. Therefore user's actions are about to import existing contact list, add social media connector or add instant messaging provider. As a complementary user's actions to each of designed entities would be to create, read, update and delete (CRUD) actions from administration point of view. During the process of an identity import is mandatory a user's support where actions as human recognition are required, because data mash or the other identity conflicts are machine irresolvable. Next actions covered administrative part of application where client behavior settings ability options are shaped by Internet connectivity which could became as off-line or on-line device mode. The off-line mode recognizes active connections to Internet and automatically synchronizes changes with back-end instance. If the device does not support background listener of network status change than the responsibility of connectivity is up to user over corresponding passive sync actions. While the active on-line mode requires requests to be served just in time and therefore personal identities would be provided any time up to date when they are required by user or by another application. Also in this case devices without support of background processing are using contact list as an provider of identities and accessibility for other application have to use embedded contact list as informational resource which is replicated upon user actions. Also there is possibility to use designed communication client where identities are automatically remotely resynchronized. Another administrative action is definition of access level permissions for each specific identity where user could globally setup public, protect or hide permission for concrete identity or in more sophisticated customization could specify permissions based on member groups.

The back-end part actions are mainly focused on background processing of connected clients or connected external identities providers. For better comprehensive overview we highlighted the entity diagram in next figure 19.2 where are required kinds of information gathered into the database. Data are consolidated within user's point of view and saved only with partial information based on social networks providers.

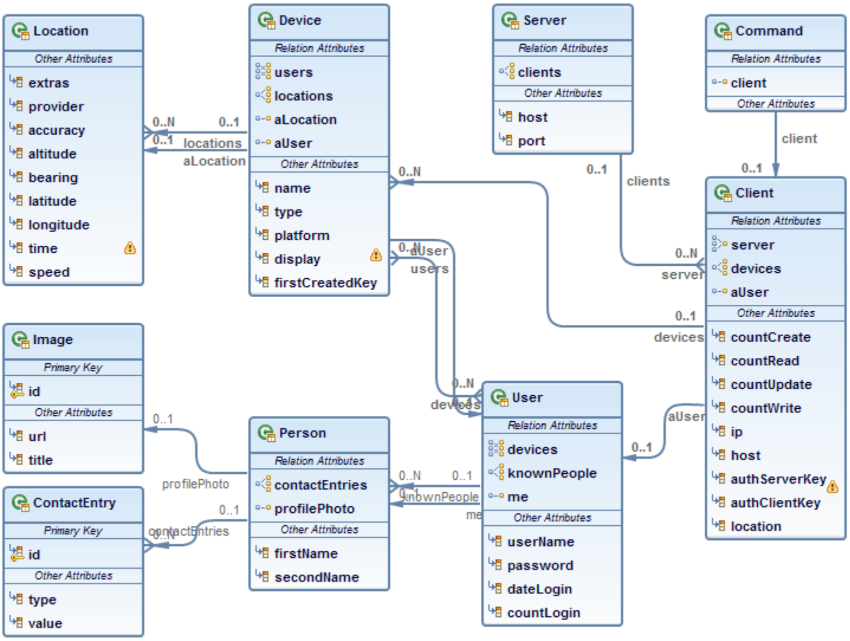


Fig. 19.2 JPA Entity diagram

The social networks and instant messengers are converging subset of external identity providers. Not all are enabling open informational exchange for independent clients. One of open exchange protocol is called Extensible Messaging and Presence Protocol (XMPP). Standardized on port 5222 and messages are exchanged over Extensive Mark-up Language (XML). We considered standard above as convenient and it will be used for interaction in further development on interface [16] with most of instant messenger providers. For social networks providers are common criteria with third side authorization process of external application which was mainly developed and enhanced by Facebook due to external social content providers who have to have only limited access to social media private data [17]. The same principles are used in G+ for accessing personal identity details.

19.5 Implementation

During the project realization we were challenging the suitability of used technologies. As the most portable solution we decided to use Java object language and supportive development framework Eclipse due to Java virtual machine (JVM) technology where clients could be implemented in any kind of device which supports embedded Java even for instance in car's radios which are able to be connected instantly to Internet [18, 19]. The prototype of testing server which provides open personal identities as a service is developed as socket Java server and running as a background process within Linux distribution (Cent OS) on virtual private server (VPS) [20]. The testing client prototype is based on Android platform because of a rapid application development (RAD) where Java is also included as a platform development language. The communication between server and client is based on message driven protocol. The messages are transferred by Java objects serialization. As server storage we used ObjectDB database engine caused by its performance results [1]. We consider that engine as the fastest in terms of usage Java Persistence API(JPA), where the Java object are annotated as database entities and therefore the transformation of any type of data between persistent Java objects in memory and physical data objects in database back and forth is automated. Currently implemented part of a concept is user interface as Android native application with touch ability. We of course plane web interface for remote management with possible device management extension therefore in the following figure 19.3 is screen of Android client application version 1 which is enabling a merge of different source of personal identities and replicates knowledge to the server.

In a certain time of period background processes are refreshing data from external resources which are announced with public accessibility. Validity for instance of email is checked with background process email checker only on untrusted inputted data.

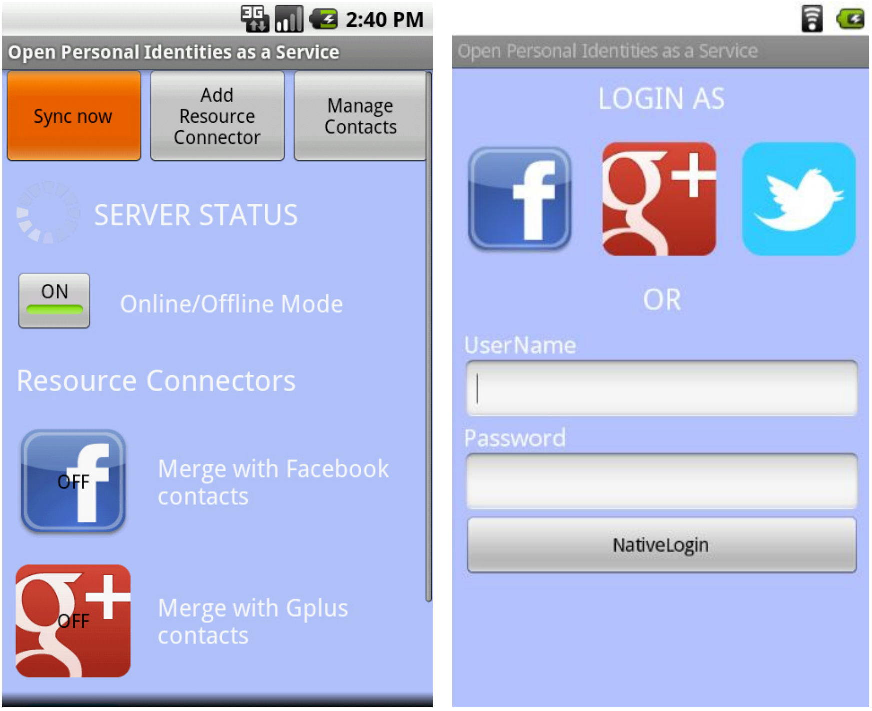


Fig. 19.3 Screens of Android client application

19.6 Conclusions

The Open Personal Identity Service solution is one part of larger scale project which covers Remote Mobile Device Management area. We consider positive influence of application usage on daily bases tasks where personal productivity increased by penetration over connected social networks. The change of any kind personal information supposed to be automatically redistributed over the connected systems. The application increase usability of maintaining social and personal identities characteristics. The open access service increase global knowledge of personal identities and positively influence human adaptability in cyber space. The real benefits of service would be recognized in further discovery with real user's behavior. The result at first step is working prototype which provides remote service of personal contact list management for mobile device users. With an increasing amount of application users the certain of personal impacts will be more obvious.

Acknowledgment. The work and the contribution were partially supported by the project (1) ESF funded project "INDOP – Innovation and Support of the Doctoral Study Program", identification number CZ.1.07/2.2.00/28.0327; (2) "SMEW – Smart Environments at Workplaces", the Grant Agency of the Czech Republic, GACR P403/10/1310; (3) specific research project "Smart Solutions in Ambient Intelligent Environments", University of Hradec Kralove under the project SP/2012/6.

References

1. JPA Performance Benchmark (JPAB), ObjectDB software Ltd. (2012), `http://www.jpab.org`
2. Ward, D.: Personal Identity, Agency and the Multiplicity Thesis. Minds and Machines 21(4), 497–515 (2011)
3. Mikulecky, P.: Remarks on Ubiquitous Intelligent Supportive Spaces. In: 15th American Conference on Applied Mathematics/International Conference on Computational and Information Science, pp. 523–528. Univ. Houston, Houston (2009)
4. Korpas, D., Halek, J.: Pulse wave variability within two short-term measurements. Biomedical papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia 150(2), 339–344 (2006) ISSN: 12138118
5. Kasik, V., Penhaker, M., Novák, V., Bridzik, R., Krawiec, J.: User Interactive Biomedical Data Web Services Application. In: Yonazi, J.J., Sedoyeka, E., Ariwa, E., El-Qawasmeh, E. (eds.) ICeND 2011. CCIS, vol. 171, pp. 223–237. Springer, Heidelberg (2011), doi:10.1007/978-3-642-22729-5_19
6. Vybiral, D., Augustynek, M., Penhaker, M.: Devices for Position Detection. Journal of Vibroengineering 13(3), 531–535 (2011)
7. Penhaker, M., Cerny, M., Martinak, L., Spisak, J., Valkova, A.: HomeCare - Smart embedded biotelemetry system. In: World Congress on Medical Physics and Biomedical Engineering, Seoul, South Korea, August 27-September 01, vol. 14, pt. 1-6, pp. 711–714 (2006)

8. Brida, P., Machaj, J., Benikovsky, J., Duha, J.: An Experimental Evaluation of AGA Algorithm for RSS Positioning in GSM Networks. Elektronika ir Elektrotechnika 8(104), 113–118 (2010) ISSN: 1392-1215
9. Chilamkurti, N., Zeadally, S., Jamalipour, S., Das, S.K.: Enabling Wireless Technologies for Green Pervasive Computing. EURASIP Journal on Wireless Communications and Networking 2009, Article ID 230912, 2 pages (2009)
10. Chilamkurti, N., Zeadally, S., Mentiplay, F.: Green Networking for Major Components of Information Communication Technology Systems. EURASIP Journal on Wireless Communications and Networking 2009, Article ID 656785, 7 pages (2009)
11. Liou, C.-Y., Cheng, W.-C.: Manifold Construction by Local Neighborhood Preservation. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 683–692. Springer, Heidelberg (2008)
12. Juszczyszyn, K., Nguyen, N.T., Kolaczek, G., Grzech, A., Pieczynska, A., Katarzyniak, R.: Agent-based approach for distributed intrusion detection system design. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 224–231. Springer, Heidelberg (2006)
13. Machacek, Z., Srovnal, V.: Automated system for data measuring and analyses from embedded systems. In: 7th WSEAS International Conference on Automatic Control, Modeling and Simulation, Prague, Czech Republic, March 13-15, pp. 43–48 (2005)
14. Bodnarova, A., Fidler, T., Gavalec, M.: Flow control in data communication networks using max-plus approach. In: 28th International Conference on Mathematical Methods in Economics, pp. 61–66 (2010)
15. Bures, V.: Conceptual Perspective of Knowledge Management. E & M Ekonomie a Management 12(2), 84–96 (2009)
16. Brad, R.: Satellite Image Enhancement by Controlled Statistical Differentiation. In: Innovations and Advances Techniques in systems, Computing Sciences and Software Engineering, International Conference on Systems, Electr. Network, December 03-12, pp. 32–36 (2007)
17. Tucnik, P.: Optimization of Automated Trading System's Interaction with Market Environment. In: Forbrig, P., Günther, H. (eds.) BIR 2010. LNBIP, vol. 64, pp. 55–61. Springer, Heidelberg (2010)
18. Thompson, T.: The Android mobile phone platform - Google's play to change the face of mobile phones. Dr. Dobbs Journal 33(9), 40-+ (2008)
19. Shih, G., Lakhani, P., Nagy, P.: Is Android or iPhone the Platform for Innovation in Imaging Informatics. Journal of Digital Imaging 23(1), 2–7 (2010), doi:10.1007/s10278-009-9242-4
20. Hii, P., Chung, W.Y.: A Comprehensive Ubiquitous Healthcare Solution on an Android (TM) Mobile Device. Sensors 11(7), 6799–6815 (2011), doi:10.3390/s110706799

Part IV
Web Systems and Network Technologies

Chapter 20
The Computer-Aided Estimate of the Text Readability on the Web Pages

Radosław Bednarski and Maria Pietruszka

Abstract. The chapter deals with the methodology of creating computer aided text readability system. Special attention concentrates on the rules helpful for a creation of readability of the text on the Web pages. Major factors are: typeface, font size, leading, text block width, text colour and background colour. Finally the system and its tests are presented.

20.1 Introduction

There are many kinds of electronic documents with text. Some of them are designed to print, others to display on the screen. Documents created by a desktop publishing software are intended for printing, then font designed for printing should be used on this one. Websites and multimedia encyclopedias, handbooks and presentations require the use of fonts dedicated to displays. This is due both to differences between the carriers and technologies of reproduction of text, as well as with differences in the way of examining the contents of the document. Note that the printed document is read and opposite to this, the display document is viewed. In addition, many Websites (e.g. blogs, home pages and pages of small businesses) are created by people who are not graphics and have no knowledge of typographic.

Existing research concerned usability [14], average reading time and text understanding [2,3,5,6,8,16], subjective feelings of the reader associated with presence and elegance of a typeface [2,4], readability of a typeface with different sizes [2,3,4,5,8,16], preferred typeface [2,3,4,6,16], colors used in text [8,9,11,13,15]. Despite many research on the text readability there is no software, which would help in the selection of typographic parameters. Existing programs only help in the

Radosław Bednarski · Maria Pietruszka
Institute of Information Technology, Lodz University of Technology, Poland
e-mail: {radoslaw.bednarski,maria.pietruszka}@p.lodz.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 211–220.
springerlink.com

selection of text color. For example Hewlett Packard Company has developed colours system verification working online. A program for colour contrast analyze in online and offline version called CCA (Colour Contrast Analyzer) was created by Jun of Wrong HTML in collaboration with Steve Faulkner of Vision Australia. Therefore, we decided to create a helpful tool for creating readable text for the Web pages for non-professional designers. We assume that besides the colour selection, it should assist in the selection of typographic parameters.

The next two sections are dedicated to typography parameters and colors of text on web pages. Then another two sections describe the proposed system to evaluate the text and present the results of tests.

20.2 Typographical Text Parameters

There are many text parameters: typeface, typestyle, font size, line length, leading (line spacing), spaces between letters (tracking), adjusting the spaces between pairs of letters (kerning) and x-height. These parameters are called typographical because of their direct impact on the layout and design text both on paper and on the screen (Fig. 20.1) [17].

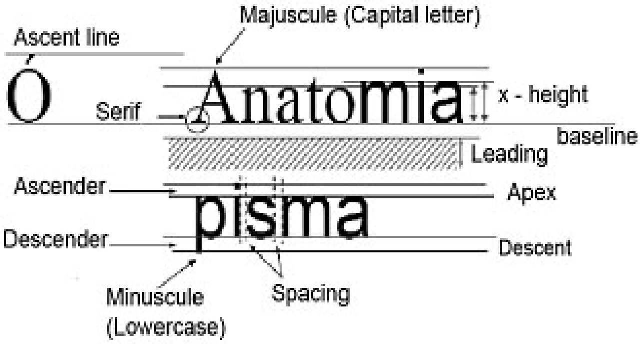


Fig. 20.1 Typographical parameters

The font is a medium that contains the information needed to reproduce typeface. The font size is a height of rectangle in which the letter has been placed. In typography a point (pt) is the smallest unit of measure of font size, but on Web pages are used pixel, em and percent too. Typeface is complete character set, designed with stylistic unity. There are three main groups of typefaces: serif (e.g. Times), sans-serif and decorative. Sheriff writing has the end lines called the sheriffs, whose task is to increase the difference in the construction of individual letters. This is preferred in most print publications, but is not recommended for screen text, particularly of small size. This is a due to differences in resolution of screen and printing: resolution of digital printers is about 600-3000 dpi while the screen resolution is about 100 dpi. A small number of pixels prevent a good representation of the letters shape, especially small items like serifs.

During planning Web page or multimedia application layout dimension of individual fields shall be set in pixels. Because there is not any correlation between the high of font and width of text block it is not easy to calculate text block with for which in one line of text there will be about 40-50 marks, according to text readability. To make it possible was created a relationship for calculating the average number of characters per line, text block width (*W*) and font size (*H*) in typographical points [1]:

$$W = 8.2H + 5.2, \text{ for 40 characters in line} \tag{20.2.1a}$$

$$W = 8.7H + 6.0, \text{ for 50 characters in line.} \tag{20.2.1b}$$

where W – width of line text [mm] and H – height of font [pt]

The basic units used in graphic design for screen is pixel. In Adobe Flash used to create web pages and web applications independent from screen resolution – 10 mm is equal to 28,5 pixels. Therefore for 1 mm accrue 2.85 pixels. Then on 1 inch = 2.54 cm accrue 72 pixels and 72 pt. Software manufacturer (Adobe) took care of the same treatment of pictures and writings in their applications. Thus, if the user gives text field width *W* in pixels it can be converted to mm by dividing by 2.58, that's means *W*[mm]=*W*[px]/2.58.

20.3 Coloured Text on the Screen

Polish standard PN-90/P-5514 "Letter printing. Characteristic of readability" recommended black or another dark print on white ground of this same intensity. This gives us maximum contrast and makes reading easier and the costs compared to colour printing are reduced significantly. The colour on the screen is free. Colorful background and/or text can be found on corporate slides multimedia presentation and Web pages [8]. Unexpectedly clear color scheme creates white text on a green background (Table 20.3.1)**.**

Man defines the colour using the three attributes:

1. Hue (*H*) is a qualitative difference of colour.
2. Saturation (*S*) is deviation from the gray colour.
3. Lightning or brightness (*Y*) indicates if the colour is closer to black or white.

Table 20.3.1 Impact of colour composition on the speed reading and the understanding of screen text; position in rank is given in the brackets [8]

No	Colours	Contrast [%]	Reading time [s]	Understanding [s]
1	Black on White	100	196 (2)	62 (1)
2	Black on Gray	60	201 (3)	59 (2)
3	White on Green	41	188 (1)	57 (4)
4	Gray on White	40	202 (4)	57 (4)
5	Yellow on Blue	77	213 (7)	58 (3)
6	Green on Yellow	30	212 (6)	55 (5)
7	Red on Green	29	209 (5)	53 (6)

Because the human eye is more sensitive for changing brightness than colour, small elements like text should be different from background brightness. Two criteria must be used to check if the selected colour is correct: brightness difference (C_B) and colour difference (C_D) [1]:

$$C_B = | Y_F - Y_B |, \quad C_D = |R_F - R_B| + |G_F - G_B| + |B_F - B_B|, \qquad (20.3.1)$$

where R, G, B are the components of the RGB colour model, and the subscripts F and B denote the text and background respectively. In the TV models [12]:

$$Y = 0.299R + 0.587G + 0.114B \qquad (20.3.2)$$

Different organizations, companies and researches recommended slightly different values of C_B and C_D (from 0 to 255 per single component of colour) [12]:

$$\mathrm{C_B} > 125 \text{ (by W3C)}, \ \mathrm{C_B} > 100 \text{ (by Ferrari [8])} \qquad (20.3.3)$$

$$C_D > 500 \text{ (by W3C)}, C_D > 400, \text{ (by HP Company).} \qquad (20.3.4)$$

We tested the 3136 colours to verify how many colours meet these criteria. It turned out that only about 11% of tested colours met both criteria (3.1) [1].

20.4 Computer-Aided System for Assessing the Readability of the Text and the Selection of Its Parameters

Existing programs are able to check text readability only in terms of text colour and background colour. As it is clear from the section 2 there are more parameters whose value can significantly affect text screen readability. The most important and able to be modified by the user are: typeface, font size, text block width, leading, text colour and background colour.

The assumption of our system is that anyone who are a graphic or a web designer, and who prepares multimedia presentation or electronic documents can be the user of the system. Certainly this system will be most useful for users who are not specialists in typography and graphic design provided that:

1. Readability degree will be determined in an understandable way for the user, for example: “the text is readable”, “the text is not readable”, “readability is low”, “readability is sufficient”, “readability is optimal”, instead of text readability = 41.3%. In this case the user does not know what to think about such evaluated text readability without additional explanation.
2. Besides the text readability evaluation, the system will generate tips which give some practical information about how to improve readability of the text. Designing of the text appearance is an individual matter, and the positive evaluation we can achieve, in different ways. Therefore, it is better not to force the user to change parameters, but only suggest parameters changing.

Basics of text evaluation are not determined precisely. The system that has such functionality should be rule-based inference system.

Linguistic variables are variables whose values are words or phrases in natural or artificial language. Ranges and corresponding values of typography linguistic variables were developed by an expert in the field of typography, professor Krzysztof Tyczkowski, for fonts designed for screen (e.g. Verdana, Tahoma, Georgia, Trebouchet, Ms Sans) and the Arial font which although designed for printing also works well on screen (Table 20.4.1).

Table 20.4.1 Linguistic variables

Linguistic variables	Range	Value of variables
Height	[6 pt, 8pt]	“Low”
Height	[9 pt, 12pt]	“Optimal”
Height	[13pt ,16pt]	“Sufficient”
Leading	[1pt]	“Low”
Leading	[2pt, 3pt]	“Sufficient”
Leading	[7pt ,9pt]	“Sufficient”
Leading	[4pt, 6pt]	“Optimal”
Number of characters	[1, 39]	“Low”
Number of characters	[40, 50]	“Optimal”
Number of characters	[51,70]	“Sufficient”
Brightnees contrast	[0, 124]	“Low”
Brightnees contrast	[125, 255]	“Optimal”
Colour contrast	[0, 400]	“Low””
Colour contrast	[401, 765]	“Optimal”

They can take the following values to specify the degree of readability: “low”, “sufficient”, and “optimal”:

1. “Low” – text is unreadable or text can be read but is not clear, font is too small, to read eyes should be closer to the monitor. There are problems to discern the characters.
2. “Sufficient” – text can be read without major problems, there are no problems with character, but reading is not comfortable – you may have problems finding another text line, sometimes eyes should be closer to the monitor.
3. “Optimal” – text can be read without any problems, there is no need to approach eyes to the monitor, there is no problem discerning the characters, reading during an extended period of time does not strain eyes.

For brightness and colour contrast two readability values were adopted: “Low” and “Optimal” which directly stems from the contrast criteria developed by W3C and HP (section 3) according to which pairs of text and background colours meet or do not meet readability condition.

Knowledge base in measurement of text readability system is rule-based. It was built based on:

- knowledge of the parameters of typography (typeface, font size, leading, the width of text box), derived from an expert in the field of graphic design and typography,
- rules deal with, text and background colors, developed by W3C consortium engaged in developed of Internet technology and HP company,
- researches and tests (see next section).

Inference module performs inference using progressive methods of reasoning, which generates new facts on the basics of existing facts and rules. This type of reasoning allows to extend the knowledge base in the future. After the user gives parameters, they are preprocessed and brought to a format acceptable by inference module. It is based on rules, assessed value of each linguistic variable, and final text readability. The final stage is displaying of partial degrees and final text readability value. System will display suggestions to improve each one of text parameters when the readability is not optimal.

The inference engine using the rules specifies the value of R_l, R_w, R_k linguistic variables. Readability is evaluated due to text height (R_h), line spacing (R_l), average number of characters per lines (width of text field) (R_w), chosen colors of text and background (R_k). They are converted to numerical values “Low – 0”, “Sufficient – 1”, “Optimal – 2” and written to array readability ratings defined as follow:

$$\text{var } \textit{readability}\text{: Array} = [R_h, R_l, R_w, R_k]; \qquad (20.4.1)$$

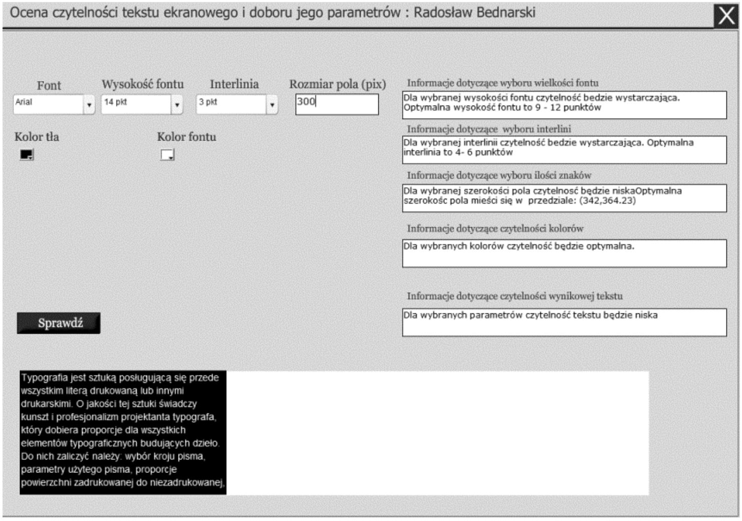
This will facilitate processing to get final text readability assessment:

$$\text{var } \textit{estimate}\text{: Array} = [0, 1, 2]; \qquad (20.4.2)$$

The final readability assessment (R) will be minimal of assessment of individual parameter. This due to the fact that it is sufficient that only one of parameter was mismatched and it will negatively affect readability of the whole text:

$$R = \min(R_h, R_l, R_w, R_k) \qquad (20.4.3)$$

The here described system for assessment text readability was programmed in Adobe ActionScript Language. The user enters name of font, font size and line spacing in typographical points, width of text field in pixels. Text and background colours are selected using Color Picker component. In the lower left corner of the screen the system displays a text sample of selected parameters. On the right side the system displays partial evaluation and the optimal value and final assessment of readability (Fig. 20.4.1).



Ocena czytelności tekstu ekranowego i doboru jego parametrów : Radosław Bednarski

Font: Arial

Wysokość fontu: 14 pkt

Interlinia: 3 pkt

Rozmiar pola (pix): 300

Kolor tła: ■

Kolor fontu: □

Informacje dotyczące wyboru wielkości fontu: Dla wybranej wysokości fontu czytelność bedzie wystarczająca. Optymalna wysokość fontu to 9 - 12 punktów

Informacje dotyczące wyboru interlini: Dla wybranej interlinii czytelność bedzie wystarczająca. Optymalna interlinia to 4- 6 punktów

Informacje dotyczące wyboru ilości znaków: Dla wybranej szerokości pola czytelnosć będzie niskaOptymalna szerokośc pola mieści się w przedziale: (342,364.23)

Informacje dotyczące czytelności kolorów: Dla wybranych kolorów czytelność będzie optymalna.

Informacje dotyczące czytelności wynikowej tekstu: Dla wybranych parametrów czytelność tekstu będzie niska

Typografia jest sztuką posługującą się przede wszystkim literą drukowaną lub innymi drukarskimi. O jakości tej sztuki świadczy kunszt i profesjonalizm projektanta typografa, który dobiera proporcje dla wszystkich elementów typograficznych budujących dzieło. Do nich zaliczyć należy: wybór kroju pisma, parametry użytego pisma, proporcje powierzchni zadrukowanej do niezadrukowanej,

Fig. 20.4.1 User interface

20.5 Tests of the System

To verify accuracy and usefulness of the system there were performed two types of test. The first test was performed by two experts in the fields of typography and graphic design. Its purpose was to check and correct the knowledge put into a system. The second test was performed by end-users with no experience in graphic design to check if the system can help them in text designing. Both tests were performed on a PC with Windows XP operating system, and a LCD monitors with the 1280x800 addressability. Assessed was a Polish text taken from a book by Krzysztof Tyczkowski “Lettera Magica”.

Expert tests were conducted on two levels: test of typographical parameters and test of colours. Both tests were developed by Krzysztof Tyczkowski, PhD – an expert who pre-consulted readability assessment of typographical parameters in a certain range of values. For test were asked two experts in field of typography and graphic design. The test of typographic parameters was based on a subjective assessment of readability of text complex one typeface (Arial) with different heights and line spacing. The examined text had a fixed number of about 40 characters per line. The evaluation of this same text has been developed by the system. Then differences were analyzed. The main difference arose for the space line parameter. Initially it was dependent on the height of the font. Expert tests have shown that it should not be done for text on the screen so a change to the system was done.

The colour test consisted of evaluation of white, grey or black text with optimal typographic parameters, showed on colored background. In total 52 pairs of colours were tested with the focus on basic colours.

Conclusions of an expert from the tests were as follows:

- most readable is black text on a white background,
- text in dark gray on a white background located in the assessment of optimal readability,
- white and gray text displayed on a black background is very well noticeable, but requires more time for reading,
- text presented on a colour background is not very active - the first note is background colour not the text,
- readability of text on a yellow or green background is good.

The second type of test was performed for end-user of the system. The test group consisted of 27 people aged between 19 and 57. Some people who participated in the study wore prescription glasses, also during the test. All subjects were experienced Internet users – have used it for a minimum of three years. Varied age of participants of the test allowed to check the influence of age on user preferences regarding the various text parameters.

The test group was divided into two parts. The first group consisted of 11 people tried to improve the parameters of the original text without supporting of the system. Participants did not see the assessment of various parameters of the text, nor the final assessment, or system suggestions for improving parameters. This test is called „test without application". The second group in which 16 people were tested saw both, the assessment of individual parameters as well as suggestions for their amendment. This test is called "test with application".

Initial test settings for the two groups were as follow: font size 8 pt, leading 2 pt, width of text fields 200 px. The colours were black for the background and dark grey for the text. In assessment of the system, readability of all parameters was determined as "Low". Therefore the readability of the entire text is also "Low". The task for both groups was to assess the initial text parameters according to their own feelings, changing the parameters of the text so that the text was more readable and evaluation of text parameters after their changing. After the evaluation by users who were not using the full system, their assessment was compared with application assessments for the parameters proposed by user. Each of the respondents in both the first and the second test group filled a questionnaire in which he answered the initial question and question related to the parameters of the text. In the test the initial text parameters was assessed by the user, then the user proposed their own settings and their assessment. Finally users of the first group were asked whether their parameters improve text readability and if there is a need to create a system that helps in setting the parameters of the text in order to improve its readability. Members of the second group were asked whether the parameters suggested by the application improve text readability and if the application designed for assessing the readability of the text and the selection of its parameters is useful for creating readable web pages.

The evidence confirming hypothesis about the usefulness of the system will show that user's readability improvements aided by the system are able to make more favorable adjustment of parameters than users who improved parameters without aid of the system. This hypothesis will be conformed thanks to a comparison of the text evaluation in the two groups that demonstrates that more correct decision occurred in the group using aid of the system.

The system will be considered as useful if following hypotheses will be proved:

1. Users in the questionnaire find that use of the evaluation system has improved the text readability.
2. In group evaluating the text with aid of the system there will be no Low readability assessment
3. In the group using the system will occur frequently more text readability improvements than in the group that does not use the aided system.
4. Text readability evaluation in group that does not use the aided system will be different from evaluation of the system.

For statistical analysis Fisher's exact test was performed. This method was chosen because of the conditions of experimentation. Fisher's exact test is performed when multiplicity sample below 40 and assuming that any expected a value is less than 5. The studies satisfy conditions of abundance and expected values. This test showed significant differences in frequency readability of text improving between group who used full version of aided system and group who used version without aid (Table 20.5.1). From the analysis for improving text parameters indicate that use of a system - aided text evaluation significantly improves the text readability.

Table 20.5.1 Summary results of the comparision of frequency of improvement of text readability; *N* – number of people in group, *f* – fraction, *p* – significance level.

	Improvement of text readability						
	Yes		No		All		
Group	N	f	N	f	N	f	p≤ 0.05
With system aid	16	1.00	0	0	16	1.00	
Without system aid	6	0.54	5	0.46	11	1.00	0.006
All	22	1.00	5	0.46	27	1.00	

20.6 Conclusions

It should be noticed that the system does not require the user to set suggested parameters but only advises how to change the parameters to obtain optimal text readability. Currently the system takes into account a few basic typographical parameters such as: typeface, font size, font size, font color, background color, leading, number of characters per line and width of text block.

Programming the system in ActionScript language allows to share it online and offline. The system was created relation to websites and multimedia applications, but the universality of assumptions and approaches allows to use it also in the design of electronic documents.

Acknowledgments. The authors thank Professor Krzysztof Tyczkowski for expertise in the field of typography.

References

[1] Bednarski, R.: The estimate of screen text readability and choosing its parameters. Doctor thesis, Łódź University of Technology (2010) (in Polish)

[2] Bernard, M., Mills, M.: So, what size and type of font should I use on my website? Usability News 2(2) (2000), http://psychology.wichita.edu/surl/usabilitynews/22/font.asp (accessed April 26, 2012)

[3] Bernard, M., Mills, M., Petersom, M., Storrer, M.A.: Comparison of popular online fonts: which is best and when. Usability News 3(2) (2001), http://www.surl.org/usabilitynews/41/onlinetext.asp (accessed April 26, 2012)

[4] Bernard, M., Lida, B., Riley, S., Hackler, T., Janzen, K.: A comparison of popular online fonts: Which size and type is best? Usability News 4(1) (2002), http://www.surl.org/usabilitynews/41/onlinetext.asp (accessed April 26, 2012)

[5] Bernard, M.L., Chaparro, B.S., Mills, M.M., Halcomb, C.G.: Comparing the effects of text size and format on the readability of computer-displayed Times New Roman and Arial text. International Journal of Human-Computer Studies 59(6), 823–835 (2003), doi:10.1016/S1071-5819(03)00121-6

[6] Boyarski, D., Neuwirth, C., Forlizzi, J., Regli, S.H.: A study of fonts designed for screen display. In: Proc. of ACM CHI 1998, Conference on Human Factors in Computing Systems, pp. 87–94 (1998), doi:10.1145/274644.274658

[7] Fellici, J.: The Complete Manual of Typography. Adobe Press (2002)

[8] Ferrari, T.G.: Legibility and readability on the World Wide Web. Buenos Aires (2000), http://bigital.com/english/files/2008/04/web_legibility_readability.pdf (accessed April 26, 2012)

[9] Garcia, M.L., Caldera, C.I.: The effect of color and typeface on the readability of online text. Computers and Industrial Engineering 31(1-2), 519–524 (1996), doi:10.1016/0360-8352(96)00189-1

[10] Hall, R.H., Hanna, P.: The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. Behaviour and Information Technology 23(3), 183–195 (2004), doi:10.1080/01449290410001669932

[11] Hill, A.L., Scharff, F.V.: Readability of websites with various foreground/background color combinations, font styles and font types. In: Proceedings of the Eleventh National Conference on Undergraduate Research, vol. II, pp. 742–746 (1997)

[12] Levkowitz, H.: Color theory and modeling for computer graphics, visualization and multi-media applications. Springer (1997)

[13] Mills, C.B., Weldon, L.J.: Reading text from computer screens. ACM Computing Surveys 19(4), 329–359 (1987), doi: 10.1145/45075.46162

[14] Nielsen, J.: Designing Web Usability. Peachpit Press (1999)

[15] Scharff, L.F.V., Ahumada Jr., A.J.: Contrast measures for predicting text readability. In: Proceedings of SPIE - The International Society for Optical Engineering 5007, pp. 463–472 (2003), http://vision.arc.nasa.gov/personnel/al/papers/03ei/03ei5007-46.pdf (accessed 26 April 2012)

[16] Tullis, T.S., Boynton, J.L., Hersh, H.: Readability of fonts in the Windows environment. In: ACM CHI 1995 Proceedings (1995), doi:10.1145/223355.223463

[17] Tyczkowski, K.: Lettera Magic., Polski Drukarz Press, Łódź (2005) (in Polish)

Chapter 21
MSALSA – A Method of Positioning Search Results in Music Information Retrieval Systems

Zygmunt Mazur and Konrad Wiklak

Abstract. The chapter describes a modification of the SALSA results ranking algorithm. An assessment of its suitability for the use in a music information retrieval system has been made. The authors propose MSALSA algorithm which uses existing connections (links) between websites and music files. By using these connections, a new method of creating a ranking for music search engine results, independent from the user query, has been developed.

21.1 Introduction

Although music information retrieval is currently a well known domain, it is still a subject of a lot of research. The popular text-based MP3 search engines are now replaced by systems that provide many different ways of creating queries, like query by humming, writing a sequence of notes or providing the melodic contour of a song.

Sound files are a specific data source. Both audio signal frequencies, stored in a sound file, and additional textual information, stored in metadata, allow performing different types of queries. The term metadata means textual information extracted from a sound file, like a song name, a composer, a name of the album etc. By using one of the music databases available online, e.g. GraceNote or MusicBrainz, the correctness of the audio metadata can be verified. Creating an identification system of text information that is stored in music files by us is a relatively

Zygmunt Mazur · Konrad Wiklak
Institute of Informatics, Wrocław University of Technology
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: `zygmunt.mazur@pwr.wroc.pl`

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 221–228.
springerlink.com

complex and difficult task. It is also not a subject of this article. However, it is possible to check, in a relatively simple way, the completeness of basic information about a song, like a title, an artist's name, an album, a year of album release, or copyrights. The correctness and completeness of data that is stored in an audio file as metadata is now a major criterion of sorting query results, returned by a music search engine. Such an approach of positioning search results in modern music search engines has significant advantages – it gives certainty that a file returned as a result contains a song corresponding to metadata which is stored in that file. However, it is not difficult to see disadvantages of using only music files themselves in the results ranking algorithm. Each music file is treated as a single entity that is not related to any other music files, documents and other contents in the Internet. Unfortunately, such an approach does not correspond to the real state, and the information about the file overall popularity in the Web is permanently lost. The popularity of a music file depends not only on the individual preferences of a search engine user, but also on the popularity of websites that host that file. The first possible approach of calculating music file popularity is the number of websites that contain at least a single outgoing link, pointing to a specified music file. However, each website has equal importance, which means that any website from the Internet is equally popular. Therefore, a better solution is to use the music website popularity based on classic link-based algorithms for text documents. Such algorithms are used by modern search engines like Google. The overall sum of weights – or the maximum weight – of websites returned by the text document ranking algorithm is a rank of the music file context.

The SALSA is a well-known ranking algorithm used in text search engines. SALSA is a mixture of the best PageRank and HITS features. For each website the authority and hub scores are calculated. SALSA can be also easily calculated by using the iterative power method. It is a good idea to apply a similar approach in music search engines. The ranking algorithms which are used now in music information retrieval systems are based mostly on self-file features, like metadata etc. and individual user preferences. However, in real world the popularity of a music file depends not only on the preferences of an individual search engine user but also on the overall popularity of websites that host this file. That kind of popularity can be calculated by basing on the connections – links – between websites. Therefore, a better solution is to use the criteria of music websites popularity based on classic links-based algorithms used by modern text search engines. It must be also assumed that overall popularity of a music file is a maximum rank value from the set of websites that contain at least one link to that specified music file. That kind of rank gives information about the music file context, and can be used in the expression as a part of the overall music file rank:

$$MusicFileRank = MetaRank + \max(ContextRank) + UserRank$$

where *MetaRank* describes the file value based on self-file information like metadata, *UserRank* describes the user preferences value and *ContextRank* are values of websites that contain at least one link to the specified music file. The *ContextRank* is calculated by using one of the link-based algorithms like HITS, PageRank or SALSA.

Another important aspect is lowering the size of the computational issue by removing from the link structure those websites that don't contain any links to the music file. The article assesses the usefulness of SALSA algorithm for the possibility of its use in music information retrieval systems. It also presents a modification and test results of the SALSA algorithm due to the specificity of music files as data – the MSALSA algorithm.

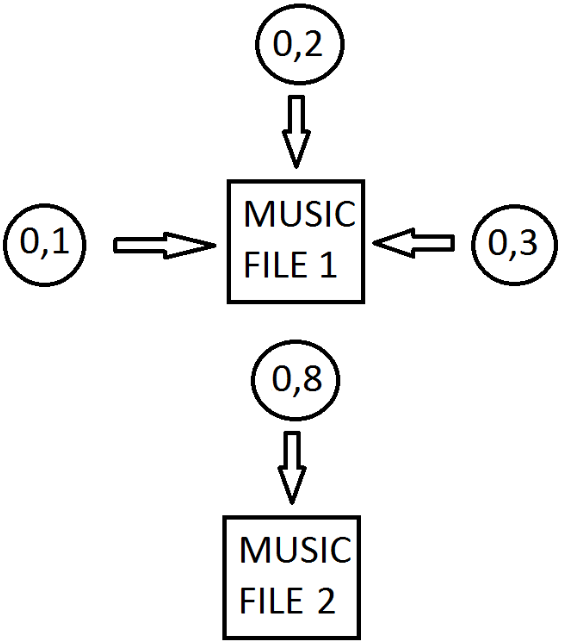


Fig. 21.1.1 The music file popularity depends on overall website popularity

21.2 Description of the MSALSA Algorithm

The SALSA algorithm is well described in [1]. Similarly as in SALSA the start point is the adjacency matrix AM. Each cell in the *i*-th row and *j*-th column of AM contains value 1 if there exists a link from the *i*-th to *j*-th website; otherwise, the cell contains 0.

From the AM the two stochastic matrices are created. First one is a row-stochastic matrix RL, and the second one is column-stochastic matrix CL. This process can be described as follows:

$$RL_{i,j} = \frac{AM_{i,j}}{k}$$

$$CL_{i,j} = \frac{AM_{i,j}}{m}$$

where k is the sum of elements in i-th row of the matrix AM and m is the sum of elements in j-th column of the matrix AM.

Next step is to determine matrices $A = CL^T RL$ and $H = RLCL^T$ where CL^T is the transposed matrix CL. After that it's necessary to remove from matrices A and H the columns and rows that contain only zeros,. Finally — after the elimination zero rows and columns — we obtain matrices *A'* and *H'* which are used in the SALSA algorithm to calculate, respectively, the authority and hub scores.

At this moment the MSALSA algorithm is the same as SALSA. However, there's a risk that matrices *A'* and *H'* can be reducible and give different result vectors — dependent on the start vector in the power method.

Irreducibility of the *A'* and H' matrices can be guaranteed by similar transformations like in the PageRank algorithm — by using the perturbation matrix and the teleportation factor. However the irreducibility can be also achieved through applying transformations on the matrices *A'* and *H'* by using the information about links from websites to music files — and that's the main point of MSALSA algorithm.

The next step of MSALSA collects information about the cardinalities of the sets of outgoing music file links from each website. For each website the probability of occurrence of the outgoing link, that is leading to the music file, is calculated and is applied to *A'* and *H*' matrices as follows:

$$MA_{i,j} = \frac{card(O_j)}{\sum_{k=1}^{a} card(O_k)} A'_{i,j}$$

$$MH_{i,j} = \frac{card(O_j)}{\sum_{k=1}^{h} card(O_k)} H'_{i,j}$$

where a and h are the number of websites respectively in the *A'* and *H'* matrices, card(O_j) is the cardinality of the set of outgoing links to music files from the j-th website. The expressions in the denominators are the sums of cardinality of the sets of outgoing links to music files from all websites respectively from the *A'* and *H'* matrix. Both *MA* and *MH* are row-stochastic matrices.

After calculation *MA* and *MH* matrices, the probability of occurrence of an outgoing link to a music file for the websites that doesn't contain any in-links – zero elements in the *MA* and *MH* matrices is calculated.

$$ZMA_i = \frac{1-\sum_{k=1}^{a} MA_{i,k}}{a}$$

$$ZMH_i = \frac{1-\sum_{k=1}^{h} MH_{i,k}}{h}$$

Please note that these “zero probabilities”, calculated separately for each row of *MA* and *MH* matrices, correspond to the perturbation matrix in the PageRank algorithm. However, it is more suitable to real world than a single fixed alpha value, and is also more suitable to music search engines, because it uses links to music files as a source of information.

The final *AMSALSA* and *HMSALSA* matrices that are used to calculate authority and hub vectors in the MSALSA algorithm are obtained by applying the *ZM* probabilities:

$$AMSALSA_{i,j} = A_{i,j} + ZMA_i$$

$$HMSALSA_{i,j} = H_{i,j} + ZMH_i$$

The next step is to apply the iterative power method to obtain the results authority and hub scores. To calculate authority score we choose stochastic initial row vector *AV* size *a*, calculation precision *delta* and perform following iterations using *AMSALSA* matrix:

```
1. $LastAV = AV$
2. $AV = AV \cdot AMSALSA$
3. IF $\| AV - LastAV \|_1 > delta$ THEN continue ELSE convergence - END.
```

Similarly, to calculate the hub score we choose stochastic initial row vector *HV* size *a*, calculation precision *delta* and perform following iterations using *HMSALSA* matrix:

```
1. $LastHV = HV$
2. $HV = HV \cdot HMSALSA$
3. IF $\| HV - LastHV \|_1 > delta$ THEN continue ELSE convergence - END.
```

21.3 Tests Results

Tests have been performed for both SALSA and MSALSA algorithms including the number of iterations in the power method that are necessary to obtain the result vector for specified fixed calculation precision and initial matrix size. The cardinality of music file links set for each website was generated randomly at the beginning of the test program. Also the connections between websites were generated randomly. To acquire high calculation precision the Java BigDecimal data type was used in implementation. Tests results are presented on the Fig. 21.3.1-21.3.4.

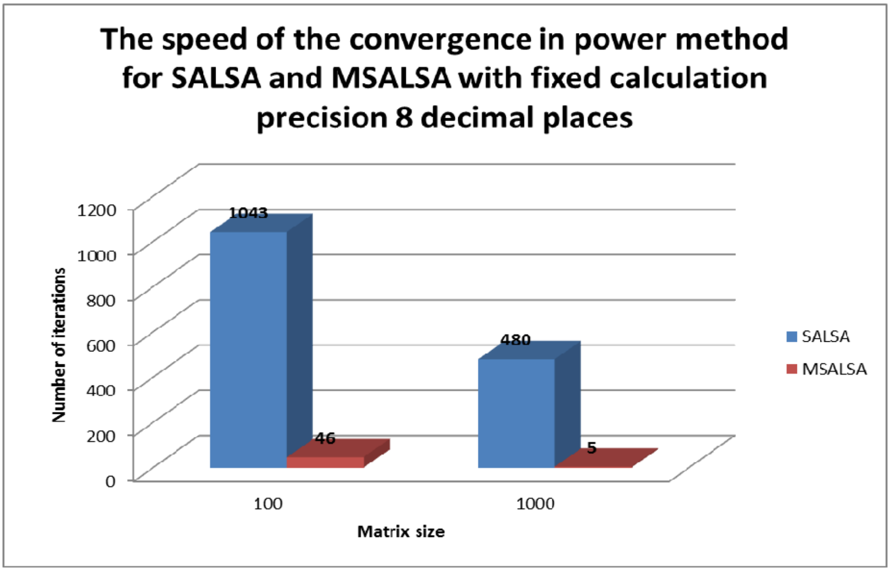


Fig. 21.3.1 The speed of the convergence in the power method for SALSA and MSALSA algorithms with fixed calculation precision of 8 decimal places

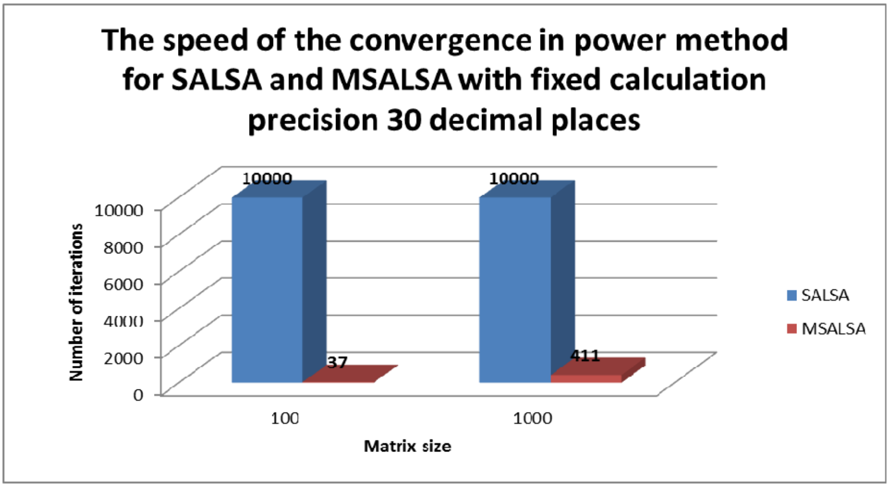


Fig. 21.3.2 The speed of the convergence in the power method for SALSA and MSALSA algorithms with fixed calculation precision of 30 decimal places

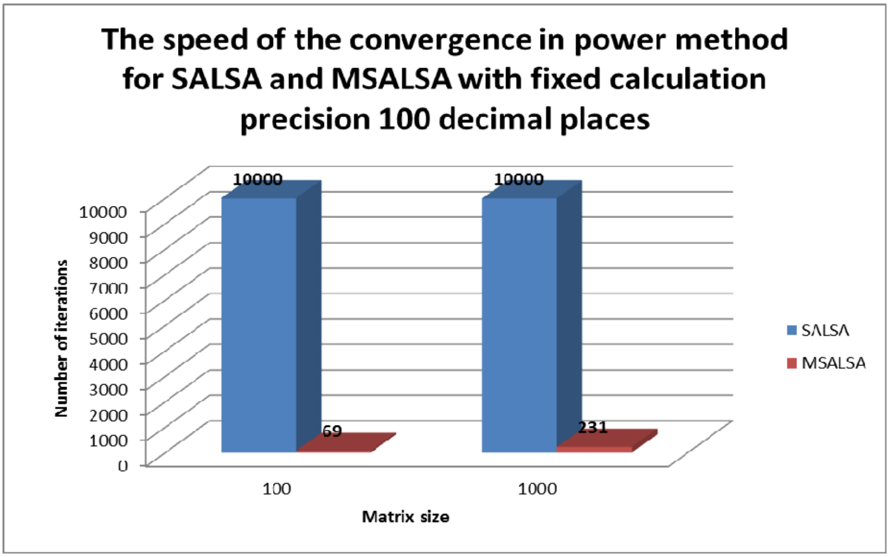


Fig. 21.3.3 The speed of the convergence in the power method for SALSA and MSALSA algorithms with fixed calculation precision of 100 decimal places

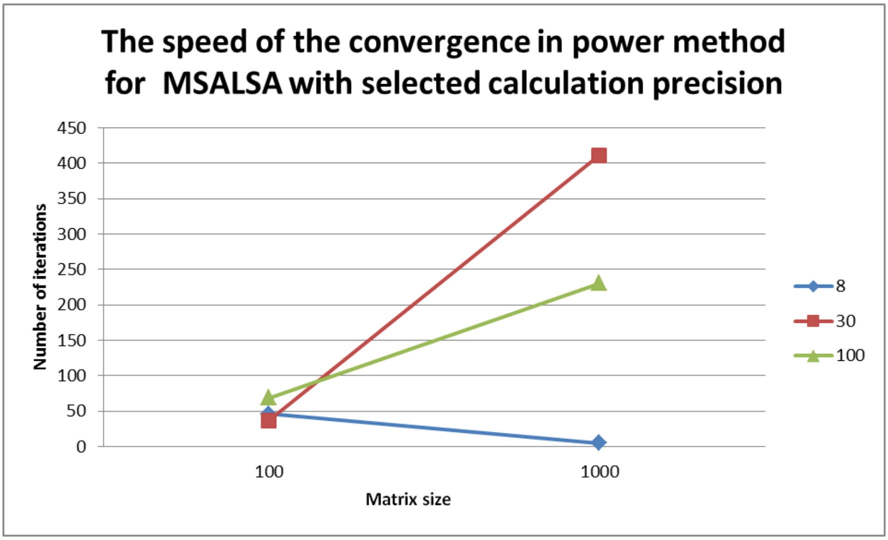


Fig. 21.3.4 The speed of the convergence in power method for MSALSA

21.4 Conclusions

The tests proved that MSALSA algorithm gives faster convergence for the authority and hub vector than SALSA. The advantage of the MSALSA algorithm is the two rank vectors – the hub and authority vector. Unlike in SALSA, the authority and hub scores are independent from the start vectors in the iterative power method. That's what makes MSALSA a more convenient ranking algorithm. It also gives more suitable results for music information retrieval scope, because information of links to music files is used. Therefore it's a good idea to introduce MSALSA as a part of a ranking algorithm in small and medium scale music search engines and music information retrieval systems. However, the MSALSA, just like SALSA, seems not to be a very scalable solution and, also is vulnerable for spamming. Therefore MSALSA may not be suitable for large scale music search engines and music information retrieval systems.

References

1. Langville, A., Meyer, C.: A survey of eigenvector methods for Web Information Retrieval. SIAM (Society for Industrial and Applied Mathematics) Review 47(1), 135–161 (2005)
2. Mazur, Z., Wiklak, K.: Music Information Retrieval on the Internet. In: Nguyen, N.T., Zgrzywa, A., Czyżewski, A. (eds.) Advances in Multimedia and Network Information System Technologies. AISC, vol. 80, pp. 229–243. Springer, Heidelberg (2010)
3. Mazur, Z., Wiklak, K.: A method of positioning search results in music information retrieval systems. In: Studia Informatica (Formerly: Zeszyty Naukowe Politechniki Śląskiej, seria Informatyka), ch. 41, pp. 527–540. Silesian University of Technology Press (2011)
4. Meyer, C.: Matrix Analysis and Applied Linear Algebra. SIAM Society for Industrial and Applied Mathematics (2004)
5. Iwanicki, A.: Zastosowanie wartości własnych macierzy (2008), `http://zaa.mimuw.edu.pl/semstud/2007-08/aiwanicki-eigenvectors.pdf` (accessed February 18, 2012)

Chapter 22
The Concept of Parametric Index for Ranked Web Service Retrieval

Adam Czyszczoń and Aleksander Zgrzywa

Abstract. Finding relevant services from a service collection which satisfy potential user query is crucial problem in Service–Oriented Computing. Parametric searching seem to be one of the basic features of Service Retrieval however there is lack of methods supporting this possibility. In this chapter we suggest a new approach to Service Retrieval of both SOAP and RESTful Web services. The usage of parametric index enables users to retrieve ranked results in accordance with specific parameters of a service. In our approach we distinguish components based on service structure which are later considered as index parameters. Because the size of such a indices is significantly big our approach uses the method of conceptual indexing which allows to reduce index size by grouping relevant service components. Additionally, for the purpose of search performance improvement for parametric index we introduce merged weight vector. Our research is supported with implementations of proposed approach by which we conducted preliminary experiments. Experimental results confirm that the proposed approach significantly reduces the index size and improves search performance of parametric service retrieval.

22.1 Introduction

Recent studies show rapid development of Service Oriented Architecture (SOA) and Web services [1]. This active progress concerns both two major classes of Web services—commonly referred to in the literature as SOAP and RESTful [5, 6, 7, 9]. Still, for service–oriented systems the key problem is finding services which satisfy information needs of potential users.

The purpose of Service Retrieval is to find relevant services from service collection satisfying given query, describing the special requirements of the user [8]. To

Adam Czyszczoń · Aleksander Zgrzywa
Wrocław University of Technology, Institute of Informatics, Wybrzeże Wyspiańskiego 27,
50-370 Wrocław, Poland
e-mail: {adam.czyszczon,aleksander.zgrzywa}@pwr.wroc.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 229–238.
springerlink.com

improve the performance of retrieving information Web services are indexed in a data structure where different index constructions allow different ways of searching. Parametric searching is one of the basic elements of Service Retrieval however currently there is lack of methods supporting this possibility.

To fill this gap, this work proposes a novel approach to index SOAP and RESTful Web Services using parametric index structure which enables ranked service retrieval. Parametric index allows user query terms to be given as parameters of a Web service, for example its name, operation, resource etc.. In case parameter is not specified all service parameters are analysed. In order to enable ranking of search results, service components are modelled in vector space. However, the problem of parametric indexing lies in huge index size. In order to reduce it our approach utilizes the method of conceptual indexing which groups relevant components into concepts. For further service retrieval performance improvement we also present merged weight vector which allows to avoid comparing each of the parameters separately to the query by index searching mechanism. In result presented approach not only reduces size of index dictionary but also improves retrieval performance.

22.2 Related Work

Basic parametric index concepts (called zone index) are presented in [4]. The concept of parametric index was also introduced and patented by Google Inc. [3]. Authors developed method of parametric indexing and parametric group processing corresponding to elements in the parametric index.

The approach of conceptual indexing was introduced by D. Peng [8] where author developed a method of indexing Web services conceptually by grouping relevant service operations into the same service concepts. In result the indexing is based on service's operations characterizes functions provided by the service instead of indexing a service with just a set of terms from the description at the entire service level. However this approach was devoted to SOAP Web services and so far there were no attempts to apply it to RESTful services as well. Additionally, presented conceptualisation approach cannot be directly applied with parametric index. The resulting service concepts govern services at operation level and information about specific service parameters cannot be extracted from resulting service concept vectors. In this chapter we use conceptual indexing to group relevant components of each service (not service collection) in order to reduce size of the index dictionary.

22.3 Web Service Structure

According to study conducted in [8] a SOAP Web service is abstracted as a set of operations where each of them represents a function provided by the service. Each service operation is a six–tuple which consists of: (i) input messages, (ii) output messages, (iii) description of operation's function, (iv) name of the operation, (v) name of the service, (vi) description of the service. Information above can be extracted from WSDL files and UDDI entries. Mentioned study considers indexing of Web services at operation level rather than service description as in traditional approaches. However for the purpose of this chapter services will be indexed in accordance with specific parameters so we cannot treat them as a set of operations.

The structure of RESTful Web service is founded on our previous study [2] based on which we consider it to be composed of: (i) name, (ii) description, (iii) version, (v) resources. The resources are in turn composed of: (vi) name, (vii) value, (viii) description, (ix) method, (x) representation. This information can be obtained from RESTful service URI and website describing the service.

The SOAP and RESTful service structural elements are similar. We can interpret the operations/resources as components of a Web service. In result the structure of each component can be described by its name, description, some input and output. When considering the I/O of SOAP component we treat them as operations with input and output messages. In the context of RESTful services we take I/O of components as resources where the input are POST, DELETE and UPDATE values and output are GET values of a resource. For precise parametric retrieval purposes we also need to keep information about component method and its representation where for SOAP services they are usually POST and XML respectively. The formal definition Web service is as follows.

Definition 1. A Web service is a quadruple of elements: $ws = \langle p_1, p_2, p_3, C\rangle$ where p_1, p_2, p_3 are *service parameters* which respectively correspond to service name, description and version. *Service parameter C* denotes service *component set* $C = \{c_1, c_2, \ldots, c_n\}$ where each component $c_i \in C$ is represented by following six–tuple $c_i = \langle A_1, A_2, A_3, A_4, A_5, A_6\rangle$ representing component elements: name, input value, output value, description, method, representation.

Above definition shows that Web services can be parametrized by its name, description, version and components which represent different functions provided by the service. Parametric searching is more precise not only because user may specify exact parameters of interest but also because services may get separate scoring for different parts. This also results in better precision than with the use of indexing at the service level or even at the operational level only.

22.4 Web Service Ranking with Vector Space Model

To enable ranking services are modelled in vector space. Each service parameter is represented as a vector in $|V|$–dimensional space where V represents index vocabulary size.

Definition 2. Web service in vector space model is represented by *service vector group* $\Gamma_{ws} = \{\overrightarrow{p_1}, \overrightarrow{p_2}, \overrightarrow{p_3}, \Gamma_C\}$ where $\overrightarrow{p_1}$, $\overrightarrow{p_2}$, $\overrightarrow{p_3}$ are *parameter vectors* of corresponding to them service parameters and Γ_C is *service component vector group* represented by $\Gamma_C = \{\overrightarrow{c_1}, \overrightarrow{c_2}, \ldots, \overrightarrow{c_n}\}$. Every $\overrightarrow{c_i} \in \Gamma_C$ stand for *component vector*. Each i-th parameter or component vector is formulated in following manner:

$$\overrightarrow{\alpha_i} = [\omega_{i1}, \omega_{i2}, \ldots, \omega_{im}] \tag{22.1}$$

where m is the vocabulary size of particular parameter or component and ω_{ij} corresponds to its weight of j-th term. Weights ω_{ij} are calculated using one of the best known combination of traditional weighting schemes in information retrieval—*tf-idf* which is defined as follows:

$$\omega_{p_i t_j} = (1 + \log \mathrm{tf}_{p_i t_j}) \cdot \log \frac{N}{\mathrm{df}_{t_j}} \tag{22.2}$$

where $\mathrm{tf}_{p_i t_j}$ is term frequency of j-th term in i-th parameter or component, N is size of parameter or component collection (separately) and df_{t_j} is the number of parameters/components in the collection that t_j occurs in.

With the above weighting scheme each parameter or component vector is represented in real–valued V–dimensional space $R^{|V|}$. Each component representation is simplified as one "bag of words" derived from its elements (component name, description, I/O etc.) which allows to represent components as above vectors.

22.4.1 Relevance Measure

In order to measure relevance between parameter or component vectors in vector space model the cosine function is used. With proposed above weighting scheme vectors have different lengths. In order to measure the relevance of vectors of different lengths they have to be normalised to the unit equal to one. In result weights have the same order of magnitude. To do that every element has to be divided by its length which is accomplished by following equation:

$$\omega_{p_i t_j} = \frac{(1 + \log \mathrm{tf}_{p_i t_j}) \cdot \log \frac{N}{\mathrm{df}_{t_j}}}{\sqrt{\sum_{m=1}^{n} ((1 + \log \mathrm{tf}_{p_i t_m}) \cdot \log \frac{N}{\mathrm{df}_{t_m}})^2}} \tag{22.3}$$

With the above equation the relevance between two length-normalised parameter/component vectors is as follows:

$$relevance(\overrightarrow{\alpha_i}, \overrightarrow{\alpha_j}) = \overrightarrow{\alpha_i} \cdot \overrightarrow{\alpha_j} = \sum_{t=1}^{m} \omega_{it}\,\omega_{jt} \tag{22.4}$$

where m is the number of terms in parameter/component vector. Equation 22.4 applies also for measuring the relevance between query and parameter or component, however query must be also length-normalised vector according to Equation 22.3.

22.5 Parametric Index

In information retrieval basic index structure is called inverted index which is composed of: dictionary – data structure that stores a collection of terms, and postings list – list of mappings to all documents in which particular term occurs. Basic concepts of parametric index structure were presented in [4] where parameters can be included in dictionary or in postings list. However extending the dictionary of parametric index by encoding parameters in terms leads to rapid increase of its size, especially for many parameters as in the case of Web services. In order to avoid index enlargement our approach is based on encoding the parameters in postings list. In parametric service retrieval we consider two situations. In the first case user specifies parameter to which his query is related to. In result the index lookup concerns only weights of specified parameter for each service that query term occurs in. For this situation parametric service retrieval is effective because the number of elements to compare with the query is equal to the size of service collection, which in this chapter is the lowest possible number of elements to check. The second situation occurs when there is no parameter specified by the user so each of the parameters is equally important. In this case, in order to find relevant services query vector must be compared against all parameter vectors for every service in the collection. In result the total number of elements necessary to check is several times bigger than service collection size which may be too big for effective service retrieval, especially if indexed services comprises many components. Our approach utilizes two methods which solve the above problem by reducing index postings list size and improving parametric index search performance. First method is based on grouping relevant service components into concepts and second method uses merged weight vector which is composed of total weights of all service parameters for each term.

22.5.1 Component Concepts

Based on the concept elaborated by D. Peng [8] relevant service components are grouped into concepts according to below definition:

Definition 3. The *component set* C_s (Definition 1) of service s is grouped into k *component concept vector group* $\Gamma_{K_s} = \{\overrightarrow{\kappa_1},\ \overrightarrow{\kappa_2}, \ldots, \overrightarrow{\kappa_k}\}$, where $\overrightarrow{\kappa_i} \in \Gamma_{K_s}$ represents a *component concept vector* of individual service s. For each $\overrightarrow{\kappa_i}$ let $G_i = \{c_1, c_2, \ldots, c_l\}$ be its *guiding set* and $\Gamma_{G_i} = \{\overrightarrow{c_1}, \overrightarrow{c_2}, \ldots, \overrightarrow{c_l}\}$ its corresponding *guiding vector group*, if for every $\overrightarrow{c_j} \in \Gamma_{G_i}$, *relevance*$(\overrightarrow{\kappa_i}, \overrightarrow{c_j}) \geq \theta$ is satisfied. The θ parameter is the *relevance threshold* between component and component concept $\overrightarrow{\kappa_i}$ represented by length-normalised concept vector and calculated as the average weight of m component terms in its guiding vector group Γ_{G_i}:

$$\overrightarrow{\kappa_i} = \frac{1}{|G_i|} \sum_{\overrightarrow{c_j} \in \Gamma_{G_i}} \overrightarrow{c_j} = \left[\frac{1}{|G_i|} \sum_{c_j \in G_i} \omega_{j1}, \frac{1}{|G_i|} \sum_{c_j \in G_i} \omega_{j2}, \ldots, \frac{1}{|G_i|} \sum_{c_j \in G_i} \omega_{jm} \right] \quad (22.5)$$

In result c service components are grouped into k component concepts where $k \leq c$ which allows to reduce the number of elements to check with the query while retrieving service.

22.5.2 Merged Weight Vector

Comparing merged weight vector to query vector reflects total service's relevance to the query. In result, while retrieving service, instead of analysing every parameter vector the searching mechanism finds relevant services based on merged weight from which it later extracts relevant parameters. In this case the number of elements to compare with the query is equal to the size of service collection, and later, for relevant parameters extraction, its equal to the number of total parameters and concepts.

Definition 4. *Merged weight vector* of Web service s is the length normalised average weight of all *service parameters*, where first three parameters represent *parameter vectors* $\overrightarrow{p_{1..3}}$ (Definition 2) of service name, description, version, and fourth parameter Γ_{κ_s} denotes service's *component concept vector group* (Definition 3) as average weight of all component concepts:

$$\overrightarrow{\sigma_s} = \frac{1}{4}\left(\sum_{i=1}^{3} \overrightarrow{p_i} + \frac{1}{|\Gamma_{K_s}|}\sum_{j=1}^{|\Gamma_{K_s}|} \overrightarrow{\kappa_j}\right) \tag{22.6}$$

In other words merged weight vector can be constructed as the sum of recursively repeated Equation 22.5 for components, then for component concepts and lastly for parameters. This not only significantly reduces index size but and increases performance of service retrieval.

22.5.3 Index Structure

According to Definition 1 we distinguished four parameters of a Web service: (i) name, (ii) description, (iii) version, (iv) components. Based on the above and referring to previously considered issues concerning service's representation in vector space model (Definition 2), component concepts (Definition 3), and merged weight vector (Definition 4) the resulting parametric index structure is as follows:

Definition 5. *Parametric index* is a $m \times n$ matrix where m denotes the dictionary size with the vocabulary composed of m terms and n represents the size of service collection with n services. For each i-th term and j-th service matrix entries a_{ij} are represented by following matrices $a_{ij} = \left[\omega_\sigma, \omega_{p_1}, \omega_{p_2}, \omega_{p_3}, \left[\omega_{\kappa_1}, \omega_{\kappa_2}, \ldots, \omega_{\kappa_k}\right]\right]$. Rows of the matrix represent *term vector groups* whereas columns represent *Web service vector groups* $\Gamma_{ws} = \{\overrightarrow{\sigma_{ws}}, \overrightarrow{p_1}, \overrightarrow{p_2}, \overrightarrow{p_3}, \Gamma_{K_{ws}}\}$, where $\overrightarrow{\sigma_{ws}}$ is represented by Equation 22.6, each *parameter vector* $\overrightarrow{p_{1..3}}$ is represented by Equation 22.1, and each *component concept vector* $\overrightarrow{\kappa_i} \in \Gamma_{K_{ws}}$ is represented by Equation 22.5.

22.5.4 Indexing Algorithm

Based on Definition 5 the following algorithm of indexing Web services is proposed:

Algorithm 1

Input: Set of Web services S. **Output:** Parametric index matrix I.

```
indexServices($S$):
1:  $I \leftarrow [\,]$
2:  for $s \in S$ do
3:     $I \leftarrow I \cup \{Null\}$
4:     $\Gamma_K, \Gamma_G \leftarrow \emptyset, \emptyset$
5:     for parameter or component as $\alpha \in s$ do
6:        $\overrightarrow{\alpha} = \left[0_0, 0_1, 0_2, \ldots, 0_{|V|}\right]$
7:        for $t \in$ uniqueTerms$(\alpha)$ do
8:           $\overrightarrow{\alpha}_t \leftarrow$getWeight$(t)$
9:        end
10:       if $\alpha$ is parameter do
11:          $I_s \leftarrow I_s \cup \overrightarrow{\alpha}$
12:       else
13:          $Rmax, \kappa pos \leftarrow$maxRelevance$(\Gamma_K, \overrightarrow{\alpha})$
14:          if $Rmax$ and $Rmax > \theta$ do
15:             $\Gamma_{G_{\kappa pos}} \leftarrow \Gamma_{G_{\kappa pos}} \cup \overrightarrow{\alpha}$
16:             $\Gamma_{K_{\kappa pos}} \leftarrow$getAverageWeight$(\Gamma_{G_{\kappa pos}})$
17:          else
18:             $\Gamma_G \leftarrow \Gamma_G \cup \{\overrightarrow{\alpha}\}$
19:             $\Gamma_K \leftarrow \Gamma_K \cup \overrightarrow{\alpha}$
20:          end
21:       end
22:    end
23:    $I_s \leftarrow I_s \cup \Gamma_K$
24:    $I_{s0} \leftarrow$getAverageWeight$(I_s)$
25: end
```

The above algorithm is based on the assumption that index dictionary is already built and its size is equal to $|V|$. Presented pseudo-code describes the process of constructing static index since no vocabulary is updated while indexing services. On the other hand, new vocabulary of newly indexed services could be appended at the end of the dictionary. The algorithm is applied for every parameter or component denoted as α of service s for which we create vector $\overrightarrow{\alpha}$ (Equation 22.1). Initially, this vector is filled with $|V|$ zeros (line 6), each zero for one term in the dictionary. In line 3, for every service we append *Web service vector group* to the I matrix. Initially this vector group has only one *Null* element to which we will append vectors. Also in the end the *Null* value will be replaced by the *merged weight vector* described in Equation 22.6. The usage of *Null* does not influence average weight calculated in the end of the algorithm. In line 4, we initialise service's *component concept vector group* Γ_K and *guiding vector group* Γ_G as empty sets. In lines 7-9 we calculate length normalised *tf-idf* weights (Equation 22.3) for every unique term that parameter/component contains. Weights are assigned to their corresponding positions in $\overrightarrow{\alpha}$ vector. For the getWeight function it is assumed that values

N and df (see Equation 22.2) are known (can be computed while constructing index dictionary).

Afterwards, if α is a parameter (line 10) we append its vector to I matrix in s position (line 11). Lines 12-22 are executed for components only. In line 13 function MAXRELEVANCE calculates the relevance between each concept in Γ_K and current component vector $\overrightarrow{\alpha}$ and returns two values. First one is the value of maximum relevance from between all concepts and current component, and the second value represents position of concept with the maximal relevance to the component in set Γ_K. Because in the beginning there are no concepts and set Γ_K is empty, function MAXRELEVANCE returns $Null$ as the value of $Rmax$, so condition in line 14 is not satisfied. Therefore, in line 18 we add new set of guiding vectors with only one element $\overrightarrow{\alpha}$ to the vector group Γ_G. Each set in Γ_G corresponds to one concept in Γ_K, therefore in line 19 we add current vector $\overrightarrow{\alpha}$ to the concept vector group Γ_K. In result we created first concept κ_0 which in the beginning is the same as first component vector $\overrightarrow{\alpha}$. Since at this point set Γ_K is not empty and $Rmax$ is not $Null$, for every subsequent component lines 15-16 apply. In line 15 we add current component vector $\overrightarrow{\alpha}$ into set of which position corresponds to the position of the most relevant concept κpos. In line 16 we calculate average weight of all component vectors which belong to the guiding set $\Gamma_{G_{\kappa pos}}$ of most relevant concept. The calculation is done in a manner similar as presented in Equation 22.5. Resulting weight is assigned to concept vector $\kappa_{\kappa pos} \in \Gamma_K$ in position κpos. In other words, we update concept vector which is most relevant to current component.

After all components are processed we append resulting concept vector group Γ_K to the I matrix (line 23). In the end (line 24) we compute the value of merged weight vector in a manner as presented in Equation 22.6, and replace it with the initially created $Null$ value placed in the beginning of service vector group. In result, we receive the *Web service vector group* Γ_{ws} as presented in Definition 5. This process is repeated for every service in the input set S.

22.6 Experimental Results

Based on the concepts and algorithm presented in this chapter we implemented Web service indexing mechanism by which we conducted initial test. The aim of the experiment was to measure the effectiveness and performance improvement of proposed approach in contrast to index without conceptualisation and merged weight vector. In order to evaluate the effectiveness we used following classical information retrieval measures: precision, recall, F-measure.

The service collection of the experiment was composed of 778 SOAP Web services with total number of 5140 components and 801 parameters collected by our implementations Web Service Crawler and Indexer. The resulting index dictionary was composed of 3499 terms. The crawler's destination host was: *xmethods.net*—a directory of publicly available Web services, used by many researchers for service retrieval benchmarks. Because we are still improving our methods of RESTful Web services identification, the experiment does not include any services of this class.

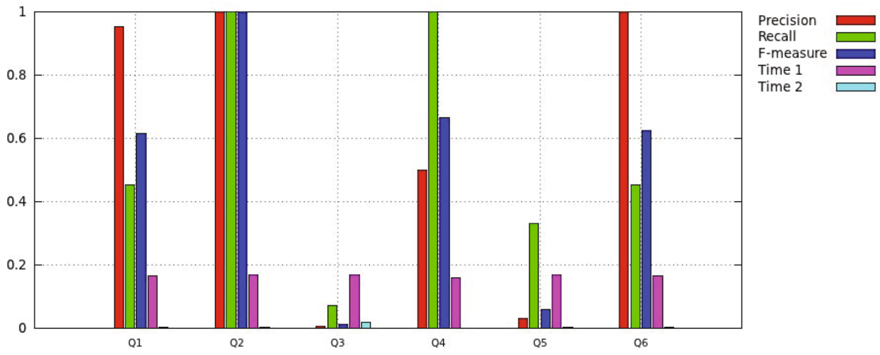


Fig. 22.1 Effectiveness of service retrieval in “basic” and “improved” index for 5 different queries.

However for the purpose of this experiment this fact does not influence our experimental results.

Size of the “improved index” based on our approach was reduced to 1815 concepts which gives 65% reduction in contrast to the total number of components. The relevance threshold θ was set to 0.65. According to [8] threshold value in range [0.5, 0.8] assures good balance between retrieval precision and performance. The experiment was carried out for following queries: *“currency exchange”*, *“weather forecast”*, *“weather forecast service”*, *“demographics”*, *“new york”* and *“send sms”*, denoted as *Q1, Q2, . . . , Q6*. The comparison of retrieval effectiveness of two index structures is presented in Fig. 22.1 where *Time 1* corresponds to response time in seconds for the first index structure and *Time 2* for the second structure.

For given queries the effectiveness of service retrieval in both structures was the same. There was however substantial difference in response time. On average it was equal to 0,167s for first structure and 0,006s for second. This means that our approach improved search response time by 97%. Despite the fact the effectiveness was the same, retrieval results differed in the ranking order. The ranking threshold was set to ten services with best relevance to the query and assumed that services which do not belong to this group are considered as irrelevant.

The effectiveness was very high for unique query terms contained by only few services. For popular terms the effectiveness was lower because among big group of retrieved services only few of them were relevant. For example for query “new york” only one service was relevant but the group of retrieved services was very big because many of them contain term “new”. Similar result can be observed for queries “weather forecast” and “weather forecast service” where therm “service” decreases the overall effectiveness. However this drawback is not significant for ranked results where the most relevant ones are returned on top.

22.7 Conclusions and Future Work

Retrieving reusable services for composing new service-oriented solutions is fundamental to Service Oriented Computing. In this chapter we presented approach which fills the gap for ranked parametrical retrieval of both SOAP and RESTful Web services. Our research includes formal definition of Web service which distinguishes basic service parameters. Secondly we presented how parametrised services can be modelled in vector space in order to enable ranking. Next, we introduced the actual structure of parametric index together with indexing algorithm. For performance measures and index size reduction purposes our approach also utilizes the method of conceptual indexing in order to group relevant service components into concepts. For further service retrieval performance improvement we also present merged weight scheme of all parameters of particular service. The resulting mechanism allows to avoid comparing weights of every parameter to given query. Experimental results show that proposed approach allows to significantly reduce index size and return search results almost 30 times faster while keeping the effectiveness of service retrieval at the same high level as in standard parametric index. In order to provide more searching possibilities for the user, in our future work a method of efficient indexing for phrase and tolerant retrieval will be developed.

References

1. Al Masri, E., Mahmoud, Q.H.: Investigating web services on the world wide web. In: The 17th International Conference on WWW, pp. 795–804. Springer, Beijing (2008)
2. Czyszczoń, A., Zgrzywa, A.: An artificial neural network approach to RESTful Web services identification, pp. 175–184. Oficyna Wydawnicza Politechniki Wrocławskiej (2011)
3. Latarche, N., Wang, J.: Apparatus and method for parametric group processing. Patent No.: US 7,461,085 B2 (2008)
4. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
5. Booth, D., et al.: Web services architecture. W3C Working Group Note, February 11, World Wide Web Consortium (2004), http://www.w3.org/TR/ws-arch/
6. Erl, T., et al.: SOA Governance: Governing Shared Services On-Premise and in the Cloud. Prentice-Hall (2011)
7. Pautasso, C., Zimmermann, O., Leymann, F.: Restful web services vs. big web services: Making the right architectural decision. In: 17th International WWW Conference, Beijing (2008)
8. Peng, D.: Automatic Conceptual Indexing of Web Services and Its Application to Service Retrieval. In: Jin, H., Rana, O.F., Pan, Y., Prasanna, V.K. (eds.) ICA3PP 2007. LNCS, vol. 4494, pp. 290–301. Springer, Heidelberg (2007)
9. Richardson, L., Ruby, S.: RESTful Web Services: Web Services fot the Real World. O'Reilly Media, Inc., Sebastopol (2007)

Chapter 23
Web-Based User Interface for SOA Systems Enhanced by Ontology

Marek Kopel and Janusz Sobecki

Abstract. In the chapter a method for ontology enhanced user interface adaptation for SOA systems is presented. SOA paradigm defines a set of methodologies for building a software out of interoperable services. Most of the services have an interactive character so they need a corresponding user interface to be defined. Today these interfaces are implemented by the software designers and programmers manually. Today, however there is a need for more flexible user interface authoring and automatic generation, which is based on the services input and output parameters and their ontology-based description.

23.1 Introduction

The problem of user interface adaptation has been addressed in many papers before, for example [12] and [15]. The user interface is usually adapted to the personal user needs and/or environment settings [10]. User interface adaptation is also one of the important problems of systems based on the SOA (Service Oriented Architecture) paradigm. SOA paradigm was developed in the beginning of a new millennium as a reaction of the software community to the discontent over interoperability, reusability, and other issues of traditional software development best practices and standards [1]. The introduction of a new service oriented best practices that are shaped to deliver strategic enterprise solutions and to overcome different shortcomings that appear at the tactical level. It is also said that SOA framework is a general guide how to conceptualize, analyze, design, and architect their service-oriented assets [1]. One of the basic features of the SOA based systems is composition of a composite service from at least several different reusable services. The consequence of this is possibly different user interface for the same service depending on the context of use of the

Marek Kopel · Janusz Sobecki
Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
e-mail: {marek.kopel,janusz.sobecki}@pwr.wroc.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 239–247.
springerlink.com

particular service. Today in most cases the user interface for the each user role and for each composite service is designed and then programmed manually by software designers and programmers. This process is usually time and money consuming, and the consumption of these resources increase with the number of different user roles and the system environments to be programmed. When we are going to design any interactive system and especially an interface for complex services we should start with the proper user model [10, 14]. We should remember However, that the ever-increasing number of different information systems also brings the increase of differences among them. These differences are caused by different reasons such as different place of living, different nationalities and cultural belongings, different generations, etc. As a result of that all these anthropological as well as cognitive differences between the users is their influence on their information needs and interaction habits. The consequence of these differences is bringing difficulties in modeling these users in the standard way [11], so more flexible solutions are necessary.

23.2 User Interfaces Enhanced by Ontology

Ontologies have been used in information systems design and development on at least several different levels, such as databases integration, business logic or Graphical User Interfaces (GUI) [13]. Application of the ontologies in the process of user interface construction is not a new idea, for example in a work [15] ontologies are applied to the problem of unification of user interface parameters in the user interface recommendation method using the ant colony metaphor. The work [13] presents an approach for mapping formal ontologies to GUI, which supports device independent GUI construction and semi automatic GUI modeling. In the work [9], however ontology-based approach of user interface development was designed for eliminating the demerits of the model-based approach, while preserving its merits, by exchange models of different interface components. In work [6] Garrett introduces elements of user experience in modelling web as a software interface or hypertext system. These elements are organized into the following layers from abstract to concrete levels: strategy, scope, structure, skeleton and surface. The representation of direct mapping from the strategy level, which is the lowest layer to the surface level which is the highest one can help users to perceive the targeted domain concept [13]. The Garrett elements can also represent UI layers in its design and development, where the ontologies deliver a framework for formal specification of the domain concepts as well as UI ones. Application of user experience elements may also reduce the user interface development expenditure. Table 1 shows mappings between levels of experience and corresponding ontological implementation together with corresponding application example using ontological framework. The example which is considered in this table shows application of personal information management system. By application of some formal ontologies we may specify the structure of interaction. The applied ontology delivers the vocabulary of any given entity concept and corresponding sub-concepts. It may also specify the domain of

Table 23.1 User Experience Elements Mappings [13]

Level of User Experience	Ontological Implementation	Example Application using Ontological Framework
Surface	Graphics Look and Feel	UIO Implementation at Graphical Library SWT, OpenGL, GTK, wxWidgets, QT
Skeleton	User Interface Ontology	Customized Textbox, List box, Selection Box, Date/Time tool, Containers, Buttons
Structure	Domain Ontology	vCard, hCard
Scope	Vocabulary (for Entities and Relations) Relations also represent Functions	Name, Address, Date of Birth, Email, Phone Number, Family Name, Zip Code
Strategy		Personal Information Management

the entity values. By defining automatic mapping functions from ontology to GUI we can deliver a tool for user interface generation.

23.3 User Interface for SOA Systems

At Wrocaw University of Technology within the itSOA grant we are developing a system called PlaTel, which is based on the SOA paradigm [2, 5, 8]. The system implements the notion of Smart Services, which is the extension of the concept of service composition [8] that enable also the execution of composed services. Smart Service, like business process is represented as a graph, where nodes represent the previously defined services and vertices connections between them and order of execution. To implement Smart Services in the working computer application a special language for their representation has been developed SSDL (Smart Service Definition Language) [8]. Smart Services described in SSDL are then interpreted and executed in Workflow Engine. SSDL is XML base language, which extends WSDL (Web Service Definition Language). The AbstractNode class is the basic class describing a node of SSDL defining composite service [8]:

```
<xs:complexType name="AbstractNode">
   <xs:sequence>
      <xs:element name="name" type="xs:string" />
      <xs:element name="nodeclass" type="xs:string" />
      <xs:element name="nodetype" type="xs:string" />
      <xs:element name="inputs" type="InputList" />
      <xs:element name="outputs" type="OutputList" />
      <xs:element name="preconditions"
                                type="PreconditionList" />
      <xs:element name="effects" type="EffectList" />
```

```
    </xs:sequence>
  </xs:complexType>
```

In the example presented above we can see that the abstract node contains information concerning required input and output parameters of services as well as elements defining preconditions and effects of a given service. The other types of SSDL nodes inherit after AbstractNode class and they can define additional requirements, which list is open. These requirements are only limited by the functionality of the Workflow Engine, however some basic types of SSDL nodes are predefined. The idea of distributed architecture poses new challenges for user interface. Although the functionality of a SOA system is implemented in a scattered manner, user who operates the system need a consistent look and intuitive interaction with the interface. One way to deal with this problem is considering any set of services that deliver a certain functionality a single, composed service. This way each user interface for a certain task is communicating with a single service endpoint, which simplifies the problem from the designers and developers point of view.

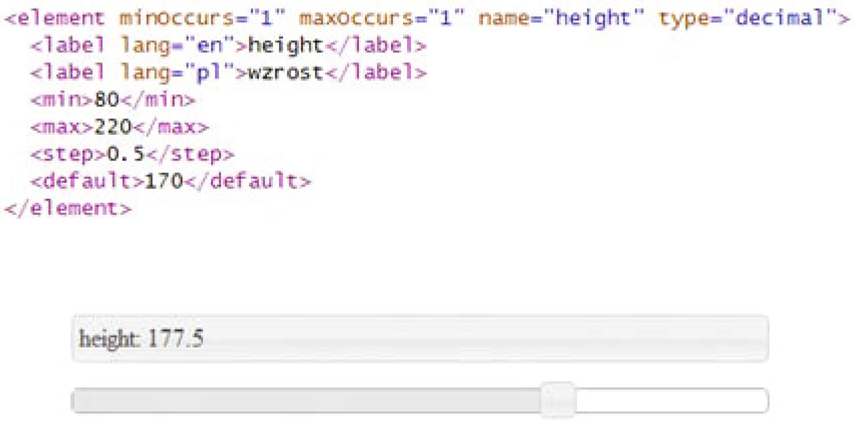
23.4 Application of Ontology Enhanced User Interface for SOA Systems

The idea of generating user interfaces for Web Services is based on the assumption that user interaction can be serialized or simplified to some kind of an HTML form. So the starting point for building a UI generator is the ability to produce a Web form. Such a Web form when submitted would deliver all the necessary inputs from user and make the interaction with a Web service possible. In order to help or even make possible for the user to fill the form some metadata are needed. In case of numerical data the metadata may inform of parameters for range and granularity of the data, i.e. min, max, step. Given the restrictions its possible to validate the data on the client side, e.g. using HTML5 Forms and setting attributes for input tags. Another application of the metadata, which also embeds validation, would be the possibility of visualizing that numerical input with a specialized widget, e.g. a slider as shown on Fig. 23.1.

In case of all the text inputs that are not free text also some validation metadata is needed. Basic validation may also be done using HTML5 Forms. They allow force entering only a valid e-mail address or other pattern matching string. But in case of more complex validation, e.g. checking if a given postal address is valid there are more sophisticated methods needed.

One approach to this problem is to restrict user input to some predefined choices. The choices are domain values defined for each non free text input, and thus turning it into a select list. Providing domain values for user inputs improves each request validation a great deal, but there are some problems with specifying the input domains:

1. Size of domain values set. If the select list in user interface contains more than 20 entries to scroll and choose from, the interface becomes user unfriendly.



```
<element minOccurs="1" maxOccurs="1" name="height" type="decimal">
  <label lang="en">height</label>
  <label lang="pl">wzrost</label>
  <min>80</min>
  <max>220</max>
  <step>0.5</step>
  <default>170</default>
</element>
```

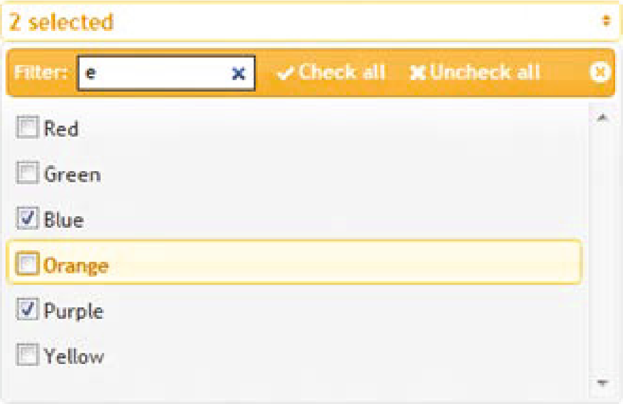
Fig. 23.1 XML formatted input validation metadata and an example visualization of the input as s slider widget.

2. Conditional nature of some inputs, i.e. when selecting in one input a certain country, than in another field allow choosing cities only from that country.
3. Multilingual/internationalization support. When designing multilingual user interfaces, the domain definitions should support multilingual values.
4. Multiple selection. Some inputs require a single value and some might need entering a specified - and greater than 1 - number of values from a defined domain.

First problem may be solved using specialized input widgets allowing filtering domain values and shrinking the select list as user enters the following characters contained in the value. An exemplary widget with that functionality is jQuery multiselect.filter shown in Fig. 23.2. Another design pattern is autocompletion, which proposes valid entries for an input based on a few characters typed in by user.

The HTML5 Specification [7] also introduces a mechanism for this design pattern: For the text, email, url, date-related, time-related, and numeric types of the input element, a new attribute list is introduced to point to a list of values that the UA should offer to the user in addition to allowing the user to pick an arbitrary value. To complement the new list attribute, a datalist element is introduced. This element has two roles: it provides a list of data values, in the form of a list of option elements, and it may be used to provide fallback content for user agents that do not support this specification.. When this standard is implemented a browser should have a native support for a long list select input, without a need for any widgets. The flip side of the domain size problem is that the number of values may be too small to visualize the input as dropdown select list. If the domain cardinality is less than 3-6, then the input is more user friendly when presented as radio buttons. Or as checkboxes, in case multiselection is needed. User interaction with a system is a flow. Enabled functionalities and inputs are dependent on the current state of the system. Projecting this fact on a UI for single Web Service it is evident that some input domains may be dependent on other entered values. Therefore there is some

```
$("select").multiselect().multiselectfilter();
```

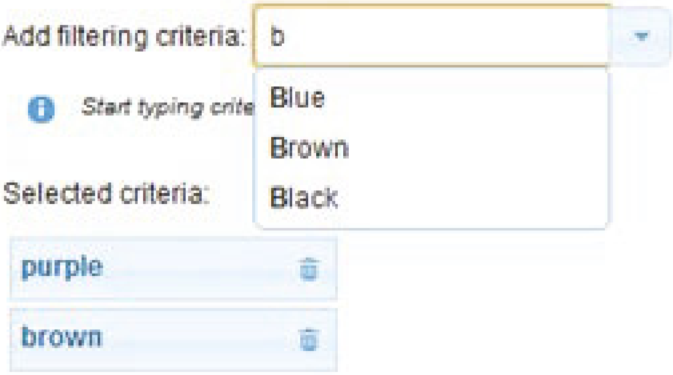


2 selected

☐ Red
☐ Green
☑ Blue
☐ Orange
☑ Purple
☐ Yellow

Fig. 23.2 jQuery multiselect.filter widget.

business logic needed to be embedded in the interface regarding user inputs. This is aspect is crucial for Web 2.0 interfaces, where adaptation of an UI based on user interaction is supposed to be instant. In order to avoid the situation where user chooses an option that is invalid in current context, a mechanism for conditional switching domain value sets is needed. For each set of domain values one should be able to define the context in which these values are valid. The context is the dependency on other inputs with values already entered or preset, but also on those not set. Other conditional aspect of input domain values is the natural language used in the interaction with a user. Although some strings, like proper names, are language independent, most of text based domains must allow defining values specific for each language. In each case of conditional domain sets, some values in one set may be context dependent while other may be independent of any UI state. In case of values being dependent on other fields values a flexible solution is needed that would allow defining complex conditions concerning many inputs and Boolean logic. Support for multilingual values are only one part of supporting a multilingual UI. For a simple text field type input a label should be language sensitive. But with specialized fields/widgets theres more to it than a label. Multilingual support extends to localization and internationalization. E.g. for date field choosing a language affects not only names of months and weeks, but a date and time format. Providing multiple selection support for a defined list of values is the harder the bigger the size of the value set gets. In case of values known to the user a simple text field with auto-complete may be sufficient. When user doesnt know the values to choose from some mechanism for filtering/shortening the list is needed. If the selected values must be presented or processed further as a sublist a design pattern presented in Fig. 23.3 may be used.



Add filtering criteria: b

Start typing crite

Blue
Brown
Black

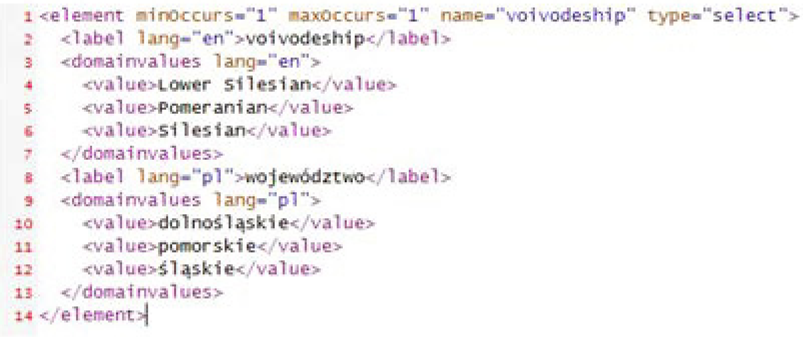
Selected criteria:

purple
brown

Fig. 23.3 Multiselection design pattern: Autocompletion combobox with a list of selected options.

23.4.1 Extending WSDL

In order to build a user interface in SOA architecture, either in automated or semi-automated way, the metadata needed for that purpose must be stored in some common format. When a UI is supposed to support interaction with a single Web Service and - as discussed earlier - each composition of services may be represented as a single, complex service, then extending the services WSDL file seems a natural decision. The extension design is based on populating WSDL nodes, which describe Web Service method parameters, with child nodes describing the metadata. An example of extending an input parameter description in WSDL is presented in Fig. 23.4. The input named voivodeship is a string type parameter. But defining it as



```
<element minOccurs="1" maxOccurs="1" name="voivodeship" type="select">
  <label lang="en">voivodeship</label>
  <domainvalues lang="en">
    <value>Lower Silesian</value>
    <value>Pomeranian</value>
    <value>Silesian</value>
  </domainvalues>
  <label lang="pl">województwo</label>
  <domainvalues lang="pl">
    <value>dolnośląskie</value>
    <value>pomorskie</value>
    <value>śląskie</value>
  </domainvalues>
</element>
```

Fig. 23.4 An extended WSDL described Web Service input parameter with label and domain values metadata.

a select with prevalidated domain values (lines 3-7, 9-13) allow visualizing it in a user interface as a widget like a singleselect version (maxOccurs=1) of those from Fig.23.2 or Fig.23.3.

Of course, in accordance with SOA paradigm, domain values need not be enlisted in the WSDL file. The ¡domainvalues¿ node can have and additional attribute specifying an URL of a corresponding set of values. This way the WSDL extension is enabled to use external, decentralized ontologies in order to get up-to-date domain value sets. The metadata are also localized for 2 languages: English (lines: 2-7) and Polish (lines: 8-13). More complex conditions for domain values can be defined in analogous way to the multilingual conditionals: for each condition a new set of values or a corresponding URL can be entered.

23.5 Discussion and Future Works

For the purpose of testing the proposed WSDL extension, a UI generator has been built. It allows generating a jQuery widget based Web interface corresponding to a given WSDL description enriched with necessary metadata. The Web interface can be generated for a specific localization (language, date format, etc.). It can be also chosen to represent pre-entered domain values as select lists or as radio buttons/checkboxes. The generator represents a proof-of-concept but only implementing the WSDL extension in a big SOA project will allow to verify the extensions usefulness and effectiveness. The generator itself may be extended to support additional widgets and visualization methods. E.g. at the moment the generator allows visualizing any geodata on a map using Google Maps API. Some instant results may also be presented in UI using HTML5 Web Forms standard output. HTML 5 specification also comes with a solution for a more crucial problem in generating or even designing user interfaces for SOA services. The problem is that UI should support the interaction in a flow of work and service calls. But since Web services are stateless, theres a mechanism for maintaining a flow, a context of a user interface is needed. For this purpose HTML5 Web Workers are proposed to be used. Thus the future work is to use Web Workers for building a complex. context sensitive user interfaces for SOA systems.

Acknowledgements. This work has been partially supported by the Polish Ministry of Science and Higher Education within the European Regional Development Fund, Grant No. POIG.01.03.01-00-008/08.

References

1. Bell, M.: Introduction to Service-Oriented Modeling. Service-Oriented Modeling: Service Analysis, Design, and Architecture, p. 3. Wiley & Sons (2008)
2. Brzostowski, K., et al.: Service discovery approach based on rough sets for SOA systems. In: Nguyen, N.T., Zgrzywa, A., Czyewski, A. (eds.) Advances in Multimedia and Network Information System Technologies, pp. 131–141. Springer, Heidelberg (2010)

3. Christensen, E., et al.: Web Services Description Language (WSDL) 1.1., http://www.w3.org/TR/wsdl
4. European Commission, From Grids to Service-Oriented Knowledge Utilities. A critical infrastructure for business and the citizen in the knowledge society (2006), ftp://ftp.cordis.europa.eu/pub/ist/docs/grids/soku-brochure_en.pdf
5. Fraś, M., Grzech, A., Juszczyszyn, K., Kołaczek, G., Kwiatkowski, J., Prusiewicz, A., Sobecki, J., Świątek, P., Wasilewski, A.: Smart Work Workbench; Integrated Tool for IT Services Planning, Management, Execution and Evaluation. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 557–571. Springer, Heidelberg (2011)
6. Garret, J.J.: Elements of user experience (April 2012), Downloaded from, http://www.jjg.net/elements/pdf/elements.pdf
7. Hickson, I.: HTML - Living Standard Last Updated 9 April 2012: Web Forms 2.0: DataList, http://www.whatwg.org/specs/web-forms/current-work/#datalist
8. Juszczyszyn, K., et al.: Service Composition in Knowledge-based SOA Systems. To be published in New Generation Computing 30(2&3) (2012)
9. Kleshchev, A., Gribowa, V.: From an Ontology-Oriented Approach Conception to User Interface Development. International Journal Information Theories & Applications 10, 87–93 (2004)
10. Kobsa, A., Koenemann, J., Pohl, W.: Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. The Knowledge Eng. Review 16(2), 111–155 (2001)
11. Kobsa, A.: Personalized Hypermedia and International Privacy. Communications of the ACM 45(5), 64–67 (2002)
12. Nguyen, N.T., Sobecki, J.: Determinantion of user interfaces in adaptive systems using a rough classification-based method. New Generation Computing 24(4), 377–402 (2006)
13. Shahzad, S.K.: Ontology-based User Interface Development: User Experience Elements Patterns. Journal of Universal Computer Science 17(7), 1078–1088 (2011)
14. Sobecki, J.: Hybrid Adaptation of Web-Based Systems User Interfaces. In: Bubak, M., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2004. LNCS, vol. 3038, pp. 505–512. Springer, Heidelberg (2004)
15. Sobecki, J.: Ant colony metaphor applied in user interface recommendation. New Generation Computing 26(3), 277–293 (2008)

Chapter 24
Simulation-Based Performance Study of e-Commerce Web Server System – Results for FIFO Scheduling

Grażyna Suchacka and Leszek Borzemski

Abstract. The chapter concerns the issue of overloaded Web server performance evaluation using a simulation-based approach. We focus on a Business-to-Consumer (B2C) environment and consider server performance both from the perspective of computer system efficiency and e-business profitability. Results of simulation experiments for the Web server system under First-In-First-Out (FIFO) scheduling are discussed. Much attention has been paid to the analysis of the impact of a limited server system capacity on business-related performance metrics.

24.1 Introduction

The issue of Web server performance evaluation has been intensively studied in recent years. The motivation has been the problem of Quality of Web Service (QoWS), resulting from transient degradation and unreliability of the Web service in the face of bursty Web traffic. To improve QoWS, many mechanisms for Web servers have been proposed, e.g. aiming at overload control [4, 11, 21] and/or request scheduling to server resources [4, 13, 17]. That research motivated development of methods and tools which make it possible to evaluate efficiency of Web servers under the control of QoWS mechanisms. Simulation methods have turned

Grażyna Suchacka
Institute of Mathematics and Informatics, Opole University, Opole, Poland
e-mail: g.suchacka@po.opole.pl

Leszek Borzemski
Institute of Informatics, Wrocław University of Technology, Wrocław, Poland
e-mail: leszek.borzemski@pwr.wroc.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 249–259.
springerlink.com

out to be useful towards that end as they are relatively inexpensive, give a possibility of detailed modeling of very complex systems, and make it possible to carry out experiments for a variety of system configurations and parameters.

A key issue in applying the simulation-based approach to Web server performance evaluation is a selection of the most adequate benchmark, i.e. a computer program emulating operation of a real Web server and collecting statistics on simulation results. A number of benchmarking tools have been developed so far, e.g. httperf [6], SPECweb99 [14], SURGE [3], S-Clients [2], WebBench [19], and WebStone [20]. However, these benchmarks have rather simplified workload model and are not oriented towards business-related performance metrics, crucial to online retailers. On the other hand, TPC-W benchmark [5] specifies an e-commerce workload, performance metrics and the system under test but it does not model details of Web server resource usage at the HTTP level. Furthermore, TPC-W implementations, e.g. [10] or [18], do not provide performance metrics related to the generated revenue and do not model various user profiles.

To partially feel the gap in this area, we have designed and developed a simulator dedicated to Business-to-Consumer (B2C) environment driven by a session-based workload generator. A workload model and a Web server system model implemented in the simulator have been characterized in [4]. The architecture of the simulator, performance metrics, and a methodology for carrying out experiments have been discussed in [15]. In this paper we discuss simulation results of experiments run with this tool for FIFO (First-In-First-Out) scheduling policy. We focus on FIFO scheduling as it is commonly applied in contemporary Web servers. The experiments have been targeted at testing characteristics of the system built according to our simulation model [4], especially at evaluating system performance in terms of various “conventional” computer system performance metrics and business-oriented metrics. Our intention was to provide a base for comparison of Web server performance under various scheduling policies, other than FIFO.

The remainder of the chapter is organized as follows. Section 24.2 characterizes workload scenarios used in simulation experiments. Section 24.3 discusses simulation results in detail. Section 24.4 concludes the chapter.

24.2 Workload Scenarios

Our simulation tool includes a workload generator which generates and transmits to the Web server system simulator a sequence of HTTP requests emulating the session-based workload (Fig. 24.1). The workload consists of two session classes: “key customers” (*KC*) acting as heavy buyers and “ordinary customers” (*OC*) acting as occasional buyers. During a single experiment, when a new session is initiated, a session class (*KC* or *OC*) is assigned to it according to a pre-specified parameter Δ_{KC}, the percentage of generated *KC* sessions in the observation window.

Each simulated customer starts their session with the home page of the online store and continues it by visiting other pages. Depending on the server system performance, each session may be finally successfully completed (when all pages in the session have been processed within a users' page latency limit T_u) or aborted (when a page response time offered by the server has exceeded T_u).

We propose five different workload scenarios differing in two main parameters characterizing Web users' behavior. The first parameter, T_u, is a threshold for users' tolerance for Web page latency, i.e. the time a user is likely to wait for a Web page being presented in a browser window. The second parameter, Δ_{KC}, means the percentage of key customer sessions arrived at the server in the observation window. Parameter values for all the workload scenarios are summarized in Tab. 24.1.

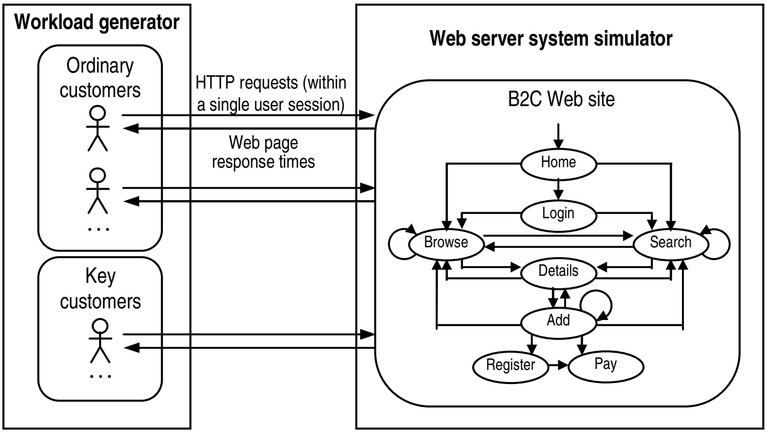


Fig. 24.1 Simulation of a user-server interaction

Table 24.1 Parameters of the workload scenarios

Workload scenario	User page latency limit T_u	Percentage of key customer sessions Δ_{KC}
Typical	8 s	10%
6s Latency	6 s	10%
4s Latency	4 s	10%
20% KC	8 s	20%
40% KC	8 s	40%

Typical workload scenario emulates the common case where T_u = 8 seconds and Δ_{KC} = 10. The value of T_u has been chosen according to the 8-second rule[1]. In

[1] Many studies for popular Web sites established a threshold of 8 seconds for the maximum latency tolerated by users [1, 16].

practice, this rule concerns the page latency perceived by a user and thereby it includes server, network and client latency components. However, we believe that such a simplification is justified in the e-commerce environment, where end-to-end latency is strongly dominated by a Web server's delay. The value of Δ_{KC} has been determined based on our preliminary simulation experiments, which showed that for our workload model 10% of generated *KC* sessions results in almost 5% of buying sessions (i.e. sessions ended with a purchase). This result is consistent with B2C session characterization studies[2].

Workload scenarios *6s Latency* and *4s Latency* differ in user page latency limits and emulate the behavior of more impatient users – T_u is equal to 6 and 4 seconds, respectively. Such situations mean that customers have bigger requirements with regard to QoWS[3].

Workload scenarios *20% KC* and *40% KC* differ in percentages of key customer sessions at the site. These scenarios mimic a B2C Web site with a large share of key customer sessions. Since the bigger number of key customers (i.e. heavy buyers) at the site leads to the bigger number of products in the shopping carts and buying sessions, these scenarios are more challenging for a Web server system.

In all the workload scenarios, the maximum number of user's retries for a given Web page request is assumed to be zero. It means that if page response time exceeds T_u, a frustrated user will give up the interaction and their session will be aborted. Such assumption leads to a little less intensive workload than for values of retries bigger than zero.

24.3 Simulation Results

Various request scheduling algorithms at the server system may be implemented in the simulation tool. In this Section a performance study for FIFO scheduling is discussed. The results may be used for a comparison of Web server performance studies for other scheduling policies to assess their relative QoWS improvements.

According to our methodology [15], each single experiment was run for a constant session arrival rate (i.e. for the constant number of new user sessions initiated per minute), for a preliminary phase duration of 10 hours and a measurement phase duration of 3 hours of the simulation time.

A group of experiments performed for the same workload and system parameters but for different session arrival rates makes up a series. Performance metrics for the whole series are presented on graphs as a function of the session arrival rate, which has varied from 20 to 300 new sessions per minute, with a step of 20. Such a range has allowed to capture a general trend of changes in observed

[2] It has been shown that the prevailing majority of online customers are only visitors who confine themselves to browsing information on products whereas the percentage of customers who end up buying something does not exceed 5% [7, 9].

[3] Some analyses for B2C Web sites indicate a 4-second page latency threshold [8, 12].

performance measures with the increase in the workload intensity, and to verify the system performance under light, medium and extremely hard workload conditions.

24.3.1 Analysis of System Performance Metrics

As the first step, we have analyzed the system throughput as the number of successfully completed user sessions per minute (Fig. 24.2). For *Typical* workload scenario (characterized by 10% of *KC* sessions and the 8-second latency for the user page latency limit), the FIFO system reaches its maximum capacity in sessions for the session arrival rate of about 100 sessions/min, and above that point the throughput continues to drop. As it can be seen, the overloaded system fails to effectively process the incoming HTTP traffic.

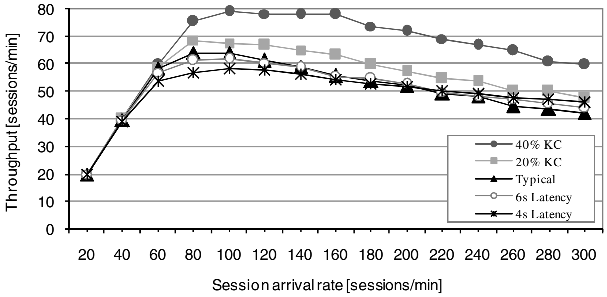


Fig. 24.2 Completed sessions per minute for different workload scenarios

It's worth observing that although above some threshold of the session arrival rate system throughput in sessions decreases with the increasing load (Fig. 24.2), throughput in HTTP requests still increases throughout the whole load range, till the maximum load (Fig. 24.3). This observation is consistent with the literature results, which have shown that considering the server capacity only at the request level may be misleading and it does not give a complete picture of the system performance. That is why we decided to use the number of successfully completed sessions per minute as the main system performance measure, instead of the number of completed HTTP requests per minute. In the following part of the chapter, "throughput" means just the number of successfully completed user sessions per minute.

For the workload scenarios other than *Typical* the maximum system capacity in sessions may slightly differ, but its decreasing tendency above some load level is clearly visible as well. The difference in the capacity levels for different workload scenarios for the same session arrival rates lies in the fact that the generated workload is a bit different depending on T_u and Δ_{KC}.

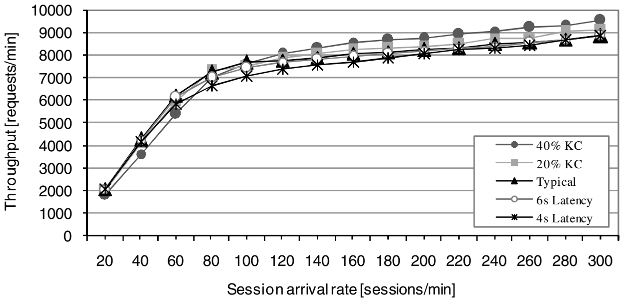


Fig. 24.3 Completed HTTP requests per minute for different workload scenarios

On the one hand, system performance depends on the user page latency limit, T_u, i.e. the maximum page response time which users are likely to tolerate. For the session arrival rate of up to 160–180 sessions/min one can observe that the more impatient users are, the lower the system throughput in sessions is, because more frustrated users give up and abort their sessions. However, for extremely heavy loads, i.e. for the session arrival rate of above 200 sessions/min, the system throughput in sessions is paradoxically higher for lower values of T_u. Our interpretation for this result is that if many sessions are aborted early, it allows one to relieve the system load and thereby more sessions in progress may be successfully completed.

On the other hand, the system performance strongly depends on the share of arrived key customer sessions, Δ_{KC}. The more *KC* sessions are, the bigger number of all completed sessions in the observation window. The reason lies in different characteristics of a session of each class [4]. First, a *KC* session is twice as short as an *OC* one, and it is a well-known fact that an overloaded Web server system discriminates against longer sessions. Second, key customers have different navigation patterns at the B2C Web site than the ordinary ones and both session classes are characterized by different sets of transition probabilities between session states. What is most relevant to our results is that key customers are much less willing to perform complex search operations at the site, which are very time-consuming. That is why *KC* sessions have a much bigger chance of a successful completion at the Web server system and in the case of workload scenarios *20% KC* and *40% KC* capacity levels are significantly higher than for the scenario *Typical*.

For the workload scenarios with a varied share of arrived *KC* sessions (*40% KC*, *20% KC* and *Typical*), the system capacity is closely related with page response times in the system. Comparing Fig. 24.2 and Fig. 24.4 one can observe for these three workload scenarios for a given session arrival rate that the higher system capacity corresponds to the lower 90-percentile of page response time (as well as to the lower median and mean values of page response times, in fact). Thus, for bigger shares of key customers interacting with the e-commerce site in the observation window the FIFO system achieves not only a bigger percentage of successfully completed user sessions but also lower page response times. On the contrary, there is no such relationship in the case of the workload scenarios differing in the user page latency limit (*Typical*, *6s Latency* and *4s Latency*). Fig. 24.4 shows that apart from the load intensity, the lower user page latency limit always implies the lower 90-percentile of page response time.

Since we are especially interested in the impact of low QoWS on e-business profitability, we further analyze revenue-oriented system performance measures.

24.3.2 Analysis of Business-Oriented Metrics

First, let us analyze an amount of *potential revenue* in a given observation window. It is defined as the total amount of money (dollars) corresponding to the total value of products that have been added to shopping carts of monitored sessions (either successfully completed or aborted). We refer to that money as "potential" revenue, because only a part of the users with full shopping carts would decide to buy something. Potential revenue obviously depends on system performance because the more users have a chance of browsing information on products and adding the selected items to shopping carts, the bigger the resulting potential revenue is. For each workload scenario the potential revenue in a 3-hour observation window increases with the increasing load until some point. Then it starts to decrease when more and more sessions are aborted at early stages and thereby, less and less users have a chance of adding products to their carts (Fig. 24.5).

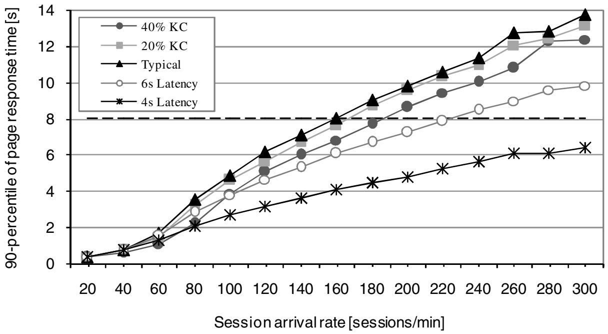


Fig. 24.4 90-percentile of page response time for different workload scenarios

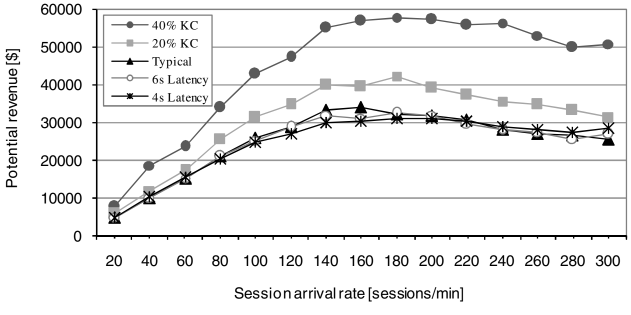


Fig. 24.5 Potential revenue in the observation window for different workload scenarios

Having analyzed the potential revenue, we evaluate achieved revenue, lost potential revenue and the percentage of achieved potential revenue for the FIFO Web server system.

The *achieved revenue* (or simply the *revenue*) is calculated as the amount of money corresponding to the total value of products in shopping carts of monitored sessions ended with purchases (i.e. the monitored *buying sessions*). Therefore, it means the actual sum of dollars earned by the online retailer during the monitored time period. As expected, for all the workload scenarios the number of successfully completed user sessions directly affects the amount of achieved revenue. Fig. 24.6 presents the revenue throughput, i.e. how many dollars have been obtained per minute through the successfully completed buying sessions. For lower system loads, the revenue throughput grows with the increasing number of users interacting with the B2C site. However, above the point of the maximum system capacity in sessions (corresponding to the session arrival rate of 80–100 sessions/min depending on a workload scenario), the revenue rate drops. A comparison with Fig. 24.2 shows that the drop of the revenue rate is even more rapid than the corresponding drop of the number of successfully completed user sessions per minute.

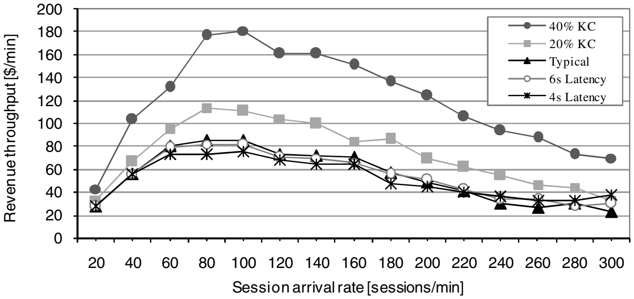


Fig. 24.6 Revenue throughput for different workload scenarios

A measure opposite to the revenue is *lost potential revenue*. It is calculated as the amount of money corresponding to the total value of products in shopping carts of monitored sessions that had been aborted due to poor QoWS, i.e. as a result of the insufficient capacity of the Web server system. Fig. 24.7 shows how much potential revenue has been lost per minute. Comparing Fig. 24.6 and Fig. 24.7 one can observe some regularity amongst the workload scenarios under overload: the higher the revenue rate, the higher the potential revenue losses. For example, when the system receives 300 new sessions per minute, the revenue of $70/min is generated in the case of scenario *40% KC* and only $23/min in the case of scenario *Typical*, while the corresponding revenue losses per minute amount to $212/min and $120/min, respectively.

In order to relate amounts of money that have been obtained and lost at various load intensity levels, the *percentage of achieved potential revenue* has been calculated. This measure gives the information on how effectively the system has processed sessions with goods in shopping carts and what percent of the potential revenue has been turned into the actual revenue. The percentage of achieved potential revenue is presented in Fig. 24.8. A value of this metric decreases with the increase in the load, indicating a bigger and bigger number of aborted sessions with goods in carts. Depending on a workload scenario, for the maximum system capacity in sessions (i.e. at the session arrival rate of 80–100 sessions/min), only 54%–80% of potential revenue has been turned into actual revenue. For the maximum session arrival rate the percentage of achieved potential revenue has been equal to merely 16%–24%.

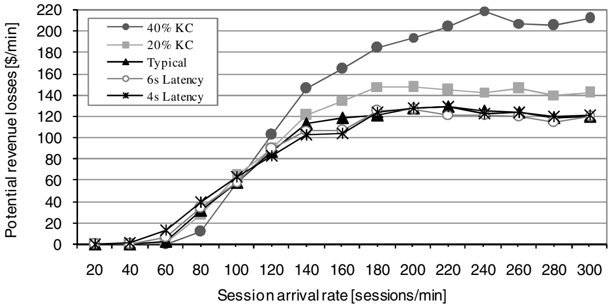


Fig. 24.7 Potential revenue lost per minute for different workload scenarios

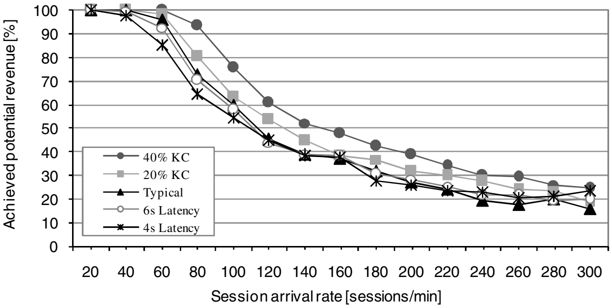


Fig. 24.8 Percentage of achieved potential revenue for different workload scenarios

24.4 Concluding Remarks

In the chapter results of the simulation-based performance study of a B2C Web server system have been discussed. The results show that FIFO scheduling fails to effectively handle the incoming peak traffic, like in real Web servers under overload. The figures presenting the system performance from the business perspective show how damaging to e-business the system's inability to effectively cope with the peak traffic can be. By not taking a negative impact of low QoWS on the number of dissatisfied customers and the amount of achieved revenue, an online retailer can miss a chance to retain the loyalty of the most valued customers and to increase revenue obtained from e-business. The results also suggest that business-oriented performance metrics are more relevant to B2C Web sites than conventional system performance metrics.

Results of the performance study for FIFO scheduling may be used as a reference point to compare the B2C Web server system performance for other scheduling policies applied at the server system. Future work will include the extension of experiments so that statistical analysis of results could be made.

References

1. 8 seconds to capture attention, Silverpop's Landing Page Report (June 2007), http://www.silverpop.com/practices/studies/landing_page
2. Banga, G., Druschel, P.: Measuring the capacity of a Web server. In: Proc. of USITS 1997, Berkeley, CA, pp. 61–71 (1997)
3. Barford, P., Crovella, M.: A performance evaluation of hyper text transfer protocols. In: Proc. of ACM SIGMETRICS 1999, Atlanta, pp. 188–197 (1999)
4. Borzemski, L., Suchacka, G.: Business-oriented admission control and request scheduling for e-commerce websites. Cybernetics and Systems 41(8), 592–609 (2010)
5. García, D.F., García, J.: TPC-W e-commerce benchmark evaluation. IEEE Computer 36(2), 42–48 (2003)
6. Jin, T., Mosberger, D.: httperf: A tool for measuring Web server performance. In: Proc. of ACM WISP, pp. 59–67. Madison, WI (1998)
7. Measure twice, cut once – metrics for online retailers, Buystream, E-Metric Research Group, http://www.techexchange.com/thelibrary/online_retail_metrics.html
8. Menascé, D.A., Almeida, V.A.F.: Capacity planning for Web services: metrics. Prentice-Hall, New York (2002)
9. Nielsen, J.: Why people shop on the Web (February 1999), http://www.useit.com/alertbox/990207.html (updated: April 2002)
10. PHARM, University of Wisconsin – Madison, http://mitglied.lycos.de/jankiefer/tpcw/index.html (access date: June 4, 2012)
11. Qin, W., Wang, Q.: An LPV approximation for admission control of an Internet Web server: identification and control. Control Engineering Practice 15(12), 1457–1467 (2007)
12. Retail Web site performance: consumer reaction to a poor online shopping experience. Jupiter Research and Akamai Report (2006), http://www.akamai.com/dl/reports/Site_Abandonment_Final_Report.pdf
13. Schroeder, B., Harchol-Balter, M.: Web servers under overload: how scheduling can help. ACM Transactions on Internet Technology (TOIT) 6(1), 20–52 (2006)
14. SpecWeb99. The standard performance evaluation corporation, http://www.spec.org (access date: April 22, 2010)

15. Suchacka, G., Borzemski, L.: Simulation-based performance study of e-commerce Web server system - methodology and metrics. In: Information Systems Architecture and Technology – Web Information Systems Engineering, Knowledge Discovery and Hybrid Computing, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, pp. 25–35 (2011)
16. The need for speed II, Zona Market Bulletin, No. 5 (2001)
17. Totok, A., Karamcheti, V.: RDRP: Reward-Driven Request Prioritization for e-commerce Web sites. Electronic Commerce Research and Applications 9, 549–561 (2010)
18. TPC-W-NYU, New York University, http://cs1.cs.nyu.edu/totok/professional/software/tpcw/tpcw.html (access date: June 4, 2012)
19. WebBench 5.0, http://cs.uccs.edu/~cs526/webbench/webbench.htm (access date: April 22, 2010),
20. WebStone - The benchmark for Web servers, http://www.mindcraft.com/benchmarks/webstone (access date: April 22, 2010)
21. Yue, C., Wang, H.: Profit-aware overload protection in e-commerce Web sites. Journal of Network and Computer Applications 32(2), 347–356 (2009)

Chapter 25
Domain Dependent Product Feature and Opinion Extraction Based on E-Commerce Websites

Bartomiej Twardowski and Piotr Gawrysiak

Abstract. The rapid growth of the Internet and social web communities has changed on-line merchandising. Opinions expressed on websites by the customers became useful information for new customers and product manufacturers. Opinion mining techniques started to be attractive as a method for processing user generated content with sentiment payload. Presented approach uses product reviews from e-commerce websites for the product feature opinion mining task. Manual data annotation process is avoided by fully automated building training data corpus. As a classifier *CRF* model is employed. Proof of concept on Polish e-commerce website was performed. Experiment has shown promising results.

25.1 Introduction

The rapid growth of the Internet and social web communities has changed on-line merchandising. Easiness of sharing information over the Internet changed the way how customers make their purchasing decisions. Before buying, customers assess and compare product on the Internet. Decisions are made based on other customers experience and opinions. This started to be a common practice, not only for costly transaction, but even for non-pricey product for everyday use. From the perspective of product creators and sellers, the on-line opinions exchange gives opportunity to collect user feedback and improves their services. Due to increasing importance of product reviews on the Internet, marketing model of many merchandising companies is changing. Building good on-line reputation become one of the primary tasks.

For potential new customers manual search for useful information from user reviews started to be tedious work. In fast growing on-line community, users generate vast amount of data everyday. Collecting what is important became difficult.

Bartomiej Twardowski · Piotr Gawrysiak
Institute of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, Poland
e-mail: B.Twardowski@ii.pw.edu.pl, P.Gawrysiak@ii.pw.edu.pl

A. Zgrzywa et al. (Eds.): Multimedia and Internet Systems: Theory and Practice, AISC 183, pp. 261–270.
springerlink.com

Opinion mining techniques started to be the answer for efficient way of processing customer created on-line reviews.

In this chapter, complete approach for feature opinion mining is presented. Feature opinion mining changes granularity of product opinion. Product is considered to have features, like: component or function. Thus, sentiment should be detected at this level, not for the product as one single unit. Presented framework uses e-commerce website with product reviews as an input repository of data. Collected data is transformed automatically into opinion mining corpus using a set of proposed tagging rules. From this point, many data mining methods can be applied to classify opinions from new product reviews. In our work, we use conditional random fields as classifier model.

25.2 Related Research

In recent years opinion mining has been an active research area providing many novel techniques. Simple methods finding sentiment orientation of a document based on positive and negative adjectives [15] evolved to the more sophisticated one with a better performance. To avoid building big lexicons of sentiment carrier adjectives, Hu and Liu in [5] presented in their work use of the WordNet system to enforce the task of adjectives sentiment classification. They consider synonyms of adjectives in lexicon to have the same polarity and antonyms opposite. But still, the seed lexicon have to be build manually. To overcome this drawback, a method of using Pointwise Mutual Information was presented in [16]. To check sentiment polarity of a given word, dependency between investigated word and: “excellent” and “poor” was checked using web search engine number of hits.

Work [9] introduced a concept of context in sentiment classification. They used intra-sentential and inter-sentential rules to check context respectively in casual clauses and adjacent sentences in passage. Many improvements were also presented in [3] as a holistic lexicon-based approach. Chapter presents how to deal with many language constructions: negation, ‘but’-clauses and context dependency. But still, polarity is based on a bootstrap lexicon with defined sentiment polarity. To overcome lexicon shortcomings, fully automated method of building lexicons was proposed in article [14]. It uses double propagation method to grow a set of the words determining opinions polarity.

New method for finding more adjectives for sentiment detection was presented in [4]. Based on a current adjectives lexicon, system using association rule mining is trying to find new adjectives in corpora. The scope of looking for associations is fixed window size measured in word distance. Founded new adjectives sentiment is verified by Church’s Mutual Information method. Words co-occurrence is checked is based on results of hits number from Google search engine.

Product feature level opinion mining using association rules was presented in [6]. Association rules were used to find high frequency features words. Infrequent features were extracted from reviews using adjacent noun phrases to know opinion words from lexicon. The main drawback of this method is limitation only to noun features. In [7] the same authors improved features extraction to class sequential rules (*CSR*) method. To build a representative rule set, they mined reviews of products with a common format containing pros and cons. The *CSR* method is not limited to noun phrases product features. However, still the noun phrase features are favored. Initial list of product feature for the rule mining is necessary. Thus, a domain expert knowledge is needed.

OpinionMiner system presented in [8] shows more comprehensive approach. System crawls review web pages to create a review database. Corpus is then annotated using specified tag set by human annotators. Fully annotated corpus is then used to train Lexicalized Hidden Markov Model (*LHMM*). *LHMM* is the variety of Hidden Harkov Model where an observable state is described by a paris: words and part of speech. A hidden state is: product features, function, component or opinion. Given a sequence of words and part of speech tags, the task is to find an appropriate sequence of output tags.

Recently in the feature based opinion mining Conditional Random Field (*CRF*) method was introduced [13]. It outperformed *LHMM* and rule based methods used as a baseline for comparison. Both methods [13, 8] and others use data mining classification technique[10] requires properly annotated training corpus. In our work, *CRF* probabilistic model have been chosen and automatic corpus preparation technique is presented.

In Polish language few opinion mining techniques were also proposed. One of the most interesting ones was [2, 1]. Authors in their work created a set of sentiment polarity detection rules for tool called *Spejd*. *Spejd* is a program allowing shallow parsing of Polish language. Proposed technique is a document level sentiment detection method, opposite to our, which is oriented on product features opinions. Experiment was carried by authors on the review about various product with the review mark collected from the Internet. For comparison of results a bag-of-words model with predefined lexicon technique was taken.

25.3 Proposed Approach

Motivated by the approaches [8, 13] which employs statistical data mining, we propose a complex framework with automatic data preparation phase for model training. In this section we describe in detail the system steps including: data acquisition, opinion mining corpus preparation and model training.

25.3.1 Data Acquisition

Nowadays most of the online merchants have their own review forums for customers. What is more, new websites emerge which are dedicated only to collect user opinions in one place (e.g. http://www.epinions.com/). Those e-commerce websites can be considered as main datasources for building our product and services review database.

Online reviews formats are different. In [7] three main groups of formats were introduced:

Format 1: Pros, cons and detail review. List of pros and cons is specified in short, mostly incomplete sentences. The detail review is used by the user to share more information and experiences with the others.

Format 2: Pros and cons. Like in *Format 1* user ought to specify pros and cons, but in this format all information is place there. Sentences are complete and well structured.

Format 3: Free form review. User is not obligated to any form.

In our approach *Format 1* is used, which is consider the most common on the Internet. A free form text review with pros and cons allows fully automatic building of the opinion mining corpus. Based on the observations and experiments, set of tagging algorithms is proposed in the next paragraph. To acquire the data, a common information retrieval technique like web crawling or scraping can be used.

25.3.2 Data Corpus Preparation

25.3.2.1 Preprocessing

Reviews collected in the previous step are user generated content with many spelling mistakes, inappropriate punctuation and arbitrary use of symbols. Sometimes the only rule on the forum of product website is human readability. Thus, the collected data should be considered noisy. Applying NLP tools, e.g. parsers or taggers, on such raw, unprepared data could result in low accuracy. Some data cleaning preprocessing steps are needed. In our framework two simple techniques are being used. First, it is symbol replacement by predefined regular expressions rules. This can be used to remove unwanted characters (e.g. “~”, stop words) or for text substitution (e.g. “ `:-)` ”, “ `:>`” to “`_SMILE_`”). Second data cleaning technique uses dictionary for spelling correction. This removes a common spelling mistakes made when computer keyboard is used. Moreover, many user comments on the Internet in Polish language do not have any diacritics symbols. Users omit them as a way for faster typing in computer keyboard prepared for English. In [2] adding missing diacritics result in a better performance.

After the data cleansing step, text data in pros, cons and detail review is processed by NLP tools for sentence boundary detection and POS tagging.

25.3.2.2 Corpus Tagset

In our work we assume two main entities in user reviews. *Product Feature* entity which is common for product physical component (e.g. battery for mobile phone), functions (e.g. internet) and its features (e.g. size, design). In [8] authors proposed separate product entities for components, functions and features. Their input corpus was manually annotated. Then it was used to extract product entities. Since our goal is to build corpora automatically from the e-commerce websites, a simplified model is more suitable and practical. Second entity is *Feature Opinion*. In users reviews, it will be an expression of user feelings and experiences for the specific product feature. Feature opinions is carrier of user emotions.

For the annotation purpose used tag set is presented in Table 25.1. Example of a tagged sentence is:

This(-) phone(F) is(-) small(O-P), has(-) high(O-P) resolution(F) screen(F) and(-) beautiful(O-P) design(F). Only(-) battery(F) is(-) poor(O-N).

Table 25.1 Tag set for annotating corpus entities

Tag name	Entity use and description
(F)	Product features, components, functions.
(O-P)	Opinion positive.
(O-N)	Opinion negative.
(O)	Opinion neutral or sentiment polarity cannot be determined.
(-)	Background. Other, non entities words.

25.3.2.3 Annotation Process

Annotation process is the main step of data preparation before training the model. As the outcome of this step opinion mining corpus is created. In this corpus every word of the review has assigned one of the metadata tag presented in Table 25.1. In the most of opinion mining works [13, 11, 6, 4] existence of a human created data corpus or lexicon is assumed. In our approach, corpus is created fully automatically.

Data annotation methods was selected based on a collected review data analysis. The results of the analysis can be presented in concise form of few observations:

Observation 1 : *single feature per pros and cons*

In the reviews form, where user can create list of pluses and minuses of the product, for a single bullet one feature is being used. In most cases, pros and cons are informal, brief text. Even when user is not obligated to such form and can leave full sentence opinions, short sentences are preferred.

Observation 2 : *same polarity for same feature in one review*

The same observations was made by authors of [14]. Review is a document written by singe person. It may contain opinions for many features. But for single feature in the review, user has invariable sentiment polarity, even if the same feature resides many times, i.e. pros and detail text.

Observation 3 : *detail text grounds pros and cons*

As mentioned earlier, in the format with pros, cons and detail review (*Format 1*), writer is using a long text in the detail review to support decision of pros and cons. User is trying to give the reason for his choice in more elaborate way: describing his experiences or comparison to the other products.

Annotation process starts with the product feature finding. Based on the observation 1 annotation of features begins with locating a noun/noun-phrases in pros and cons.

Founded words/phrases representing feature are detected in the detail review text (observation 3). For example, for pros *"good battery"*, in detail text user describe features in a more comprehensive way: *"This cell phone has great battery with long live."*. The feature here is *"battery"*. For annotating complex feature e.g. *"digital camera"*, where adjective + noun should be found, a more advanced techniques should be used. In [6] method based on the association rule was proposed. In our work we use *Pointwise Mutual Information* to find a potential word co-occurrence which can represent features.

After finding bootstrap set of the features, opinion words with their polarity are annotated. Process is based on the observation 2 and 3. Just like in the features case, primary words for the opinion are extracted from pros and cons, but in this case adjectives and adverbs (which are not features) are taken. In the next step the opinion words are searched in the detail text review. If opinion word is found in the text and its near neighborhood is the feature, then the opinion is tagged. Polarity of the opinion is decided based on from which set the opinion was derived: pros or cons. Opinion is tagged as positive or negative respectively.

While annotating the opinions words, there are few negation words which usually revers the opinion polarity in the sentence. Word *"not"* is one of the best example. Presence of this word changes polarity of the following opinion.

25.3.2.4 Expanding Annotated Data

Presented annotating methods are just simple rules to start annotation of the features and opinions. Considering that the based rules were applied on the reviews database, some portion of text is already tagged. Following methods to expand annotation process can use this seed data.

Context Opinions. Simple annotation rules are mostly constrained by a sentence boundary or even by a fixed word window. Using appropriate context in the annotation process address this limitation. First approach is to use *Feature Context Coherency*. It is based on observation 2 and 3 - sentiment charge for single feature in review is either positive or negative. Using this context, surrounding text can be considered as a positive or a negative opinion. Particularly, adjectives and adverbs. Thus, feature coherent context generates new opinion words. Secondly context strategy is used for handling context dependency in the sentences of detail review. Method introduced in [9] and used to build domain depended lexicons[3]. Examining a single sentence *intra-sentence context* can deal with conjunctions to annotated more opinions. Where opinion is found accompanied by appropriate conjunctions, the opinion context with it polarity can be spread across new words. For conjunctions like "and" polarity remains the same, for others, e.g. "but", it changes to the opposite. *Iner-sentence context* function outside single sentence boundary. It takes into consideration that reviewer usually writes opinions in many sentence with the same polarity.

Finding Synonyms and Antonyms. WordNet[1] is a great source of semantic and lexical relations of words in natural language. It is a domain independent system, but even though it can be used with a great success for the domain specific tasks. If an opinion word is annotated as positive in a context of one feature, all of its synonyms are also considered positive and antonyms as negative for this feature. Similar logic applies for negative opinions. Moreover, the WordNet system can be used in the feature expansion task. Using semantic similarity exchange words for the existing features can be found. Only appropriate high similarity value should be applied.

Result Propagation. In the seed data not all opinions and features are annotated. Many sentences have only one feature or opinion annotated. Result propagation is one of the most popular ways to raise coverage of an annotated text. Assuming that the seed data and the common annotation rules are available, we can use already annotated corpus to extract more opinions and features. Process of extraction is repeat over and over again, passing result from previous step as seed data to next one. Loop stops when no more opinions or features can be found. This method in [14] was named *double propagation*.

25.3.3 Model Training

Having data corpus annotated as presented in the previous section, in our approach the opinion mining task can be simplified to the problem of text labeling. In order to find new opinions, system should label new documents with tags from a Table 25.1.

One of the most popular method for labeling a sequential data is Conditional Random Field (*CRF*)[17]. *CRF* model can be represented as a undirected graph globally conditioned on X, the random variable representing observation sequence. In our case, the observation sequence is the sequence of words with assigned part of speech $X = (W,S)$. Formally for *CRF* method, undirected graph $G = (V,E)$ is defined, such there is a node $\upsilon \in V$ for each random variable Y_υ of Y. Where Y is a set of random variables corresponding to the labeling sequence - tag sequence $T = t_{1:N}$, where t is specified by Table 25.1. Edges E in graph G represents potential conditional dependency. If each random variable Y_υ obeys Markov model in G, then (X,Y) is conditional random field. In this work, like in the most of cases, simplified structure - linear chain[17] of graph G is considered. Thus, for our problem the most probable labeling sequence $T = t_1 \ldots t_n$ for input of words $W = w_1 \ldots w_n$ and POS $S = s_1 \ldots s_n$ is

$$argmax_T \prod_{i=1}^{N} p(t_i|W,S,t_{i-1}) \tag{25.1}$$

where probability p is given by equation

[1] http://wordnet.princeton.edu/

$$p(t_i|W,S,t_{i-1}) = \frac{1}{Z} exp(\sum_{j=1}^{N} \sum_{i=1}^{F} \lambda_i f_i(t_{j-1},t_j,w_{1:N},s_{1_N},j)) \tag{25.2}$$

Normalization factor Z and parameters λ_i are estimated for the training data like in [17]. *CRF* feature functions are taken from real samples - created opinion mining corpus. An example of such feature is

$$f_i(t_{j-1},t_j,w_{1:N},s_{1_N},j) = \begin{cases} 1 \text{ if } w_j = great, s_{j+1} = \text{NN } and\ t_j = \text{O-P} \\ 0 \text{ otherwise} \end{cases} \tag{25.3}$$

For this function if the current word is *"great"*(which express positive opinion) and the next word is noun, function is activated. Sample data for activating this function can be phrase: *"great battery"*.

Training of *CRF* model, when the graph and *CRF* feature functions are defined, is to find all $\lambda_{i:N}$ parameters values. In many free available libraries this optimization problem is solved by algorithm called Limited memory BFGS (*L-BFGS*).

25.4 Experiment and Results

In order to verify proposed approach the opinion mining system was implemented and empirical experiment was carried out. As a review source, one of the most popular Polish e-commerce sites[2] with product opinions was chosen. Selected website not only allows users to share opinions with others, but also aggregates reviews from the other Polish sites. To perform our experiment 72 654 reviews about mobile phones was acquired. That created text database with 390 461 sentences.

The preprocessing step removed unwanted words and characters. The spelling correction part was omitted. To perform sentence splitting and POS tagging TaKIPI-[12] tool was used. Afterwards, all data was loaded into a document database where the annotation process was performed. Process started with a simple annotation from pros and cons. To find complex product features a normalized *PMI* value for phrases containing at least one noun from pros or cons was calculated. It generated features like: *"battery life"*, *"easy of use"*, *"digital camera"*, *"user guide"*. After the base features and opinions were extracted, annotation expansion methods were applied. For a method using synonyms and antonyms to expand tagged data, Polish WordNet[3] was used. From all 390 461 sentences in the corpus 181 196 was tagged with at least one feature or opinion tag. In the end corpus 968 unique features was found, from which 518 appears only once. Unique opinions words found is 1949, from which 853 appears in only once.

The prepared corpus was used to train *CRF* model. The *CRF++*[4] implementation was used to perform *CRF* learning and testing. In the learning step following CRF features templates were used: $x[0,0], x[0,1], x[0,1]/x[-1,1], x[-1,1], x[-2,1]$,

[2] http://www.cokupic.pl

[3] http://www.plwordnet.pwr.wrod.pl

[4] http://crfpp.googlecode.com/svn/trunk/doc/index.html

$x[-1,0], x[-2,0], x[1,0], x[1,1], B$. Where on the first position is relation to current row and on the second a column index (0 - word, 1 - POS). B stands for *CRF++* bigram template.

To evaluate trained model 164 reviews were manually annotated. Using prepared test data corpus, algorithm performance was studied of how well features and opinions was tagged. For the result evaluation standard measure of precision, recall and F-Score was chosen. Overall precision was 0.87 with recall result of 0.69. This gives F-Score value equals to 0.77.

High precision confirms a good annotation quality and correct sentiment detection. This proves that used proposed feature context aware techniques perform well. A positive and negative polarity is passed correctly in the sentences. Even negation handling using simple approach of changing polarity in the following words performed well. However, low recall at 69 percent shows that many of tagged data was considered to be background-tagged wrongly. Human annotators founded more product features and opinions in the test data set. The main purpose of the annotation expansion methods is to minimize this uncovered data. From the proposed ones, annotation expansion based on the WordNet gave best results. In the number of new annotated words it was: 218k new annotated words for features and 103k new annotated words for opinions. But still, there are areas for improvements.

25.5 Conclusions and Feature Work

In this chapter the complete approach for product feature opinion mining was presented. Proposed method was evaluated on Polish e-commerce website. Results are promising. With precision of 0.87 and at recall level of 0.69 method can compete with many other art-of-state opinion mining systems. Lower recall in this case can be compensated by the fact, that data preparation process does not need any manual work or expert knowledge. What is more, websites with needed format of customer reviews are common on the Internet and freely accessible.

Moreover, experimental implementation and results show that text processing tools for Polish language are no longer a problem. Tools like TaKIPI and Polish WordNet are matured projects and can be used in text processing system without much risk.

Feature research involves improvements in automatic corpus creation process. Many fields in this area are still not covered, e.g. implicit feature tagging and pronoun resolution. Implicate features are features not explicitly mentioned in opinion sentence, but are known from the context. In example: *"This phone is heavy"*, implicit feature here is the weight. Pronoun resolution helps for better feature finding, especially in informal language like user product reviews. In data mining phase, the use of *CRF* shows promising results in opinion mining area. However, other *CRF* feature functions template should be evaluated in more detail. Change of *CRF* feature function templates arbitrary chosen in our experiment can result in better overall system performance.

References

1. Aleksander, B., Aleksander, W.: Shallow parsing in sentiment analysis of product reviews. In: Proceedings of the Workshop on Partial Parsing – Between Chunking and Deep Parsing (2008)
2. Buczyński, A., Wawer, A.: Automated classification of product review sentiments in polish. Intelligent Information Systems, 213–217 (2008) ISBN 978-83-60434-44-4
3. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: WSDM 2008: Proceedings of the International Conference on Web Search and Web Data Mining, pp. 231–240. ACM, New York (2008), http://dx.doi.org/10.1145/1341531.1341561, doi:10.1145/1341531.1341561
4. Harb, A., Plantié, M., Dray, G., Roche, M., Trousset, F., Poncelet, P.: Web opinion mining: how to extract opinions from blogs? In: Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology, CSTST 2008, pp. 211–217. ACM, New York (2008), http://doi.acm.org/10.1145/1456223.1456269
5. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD, pp. 168–177 (2004)
6. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: AAAI, pp. 755–760 (2004)
7. Hu, M., Liu, B.: Opinion Feature Extraction Using Class Sequential Rules. In: Proceedings of AAAI-CAAW 2006, The Spring Symposia on Computational Approaches to Analyzing Weblogs, Stanford, US (2006), http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-013.pdf
8. Jin, W., Ho, H.H., Srihari, R.K.: Opinionminer: a novel machine learning system for web opinion mining and extraction. In: KDD, pp. 1195–1204 (2009)
9. Kanayama, H., Nasukawa, T.: Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: EMNLP, pp. 355–363 (2006)
10. Ma, H.B., Geng, Y.B., Qiu, J.R.: Analysis of three methods for web-based opinion mining. In: ICMLC, pp. 915–919 (2011)
11. News, I., News, I., Corpora, B., Ku, L.W., Liang, Y.T., Chen, H.H.: Opinion extraction, summarization and tracking, doi:10.1.1.179.3061
12. Piasecki, M.: Polish tagger TaKIPI: Rule based construction and optimisation. Task Quarterly 11(1-2), 151–167 (2007)
13. Qi, L., Chen, L.: A linear-chain crf-based learning approach for web opinion mining. In: WISE, pp. 128–141 (2010)
14. Qiu, G., Liu, B., Bu, J., Chen, C.: Expanding domain sentiment lexicon through double propagation. In: IJCAI, pp. 1199–1204 (2009)
15. Stone, P.J., Bales, R.F., Namenwirth, J.Z., Ogilvie, D.M.: The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. Behavioral Science 7(4), 484–498 (1962), http://dx.doi.org/10.1002/bs.3830070412, doi:10.1002/bs.3830070412
16. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. CoRR cs.LG/0212032 (2002)
17. Wallach, H.M.: Conditional random fields: An introduction (2004)

Author Index