Advances in Intelligent Systems and Computing 506

Aleksander Zgrzywa
Kazimierz Choroś
Andrzej Siemiński *Editors*

Multimedia and Network Information Systems

Proceedings of the 10th International Conference MISSI 2016

Springer

Advances in Intelligent Systems and Computing

Volume 506

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

More information about this series at http://www.springer.com/series/11156

Aleksander Zgrzywa · Kazimierz Choroś
Andrzej Siemiński
Editors

Multimedia and Network Information Systems

Proceedings of the 10th International Conference MISSI 2016

Springer

Editors
Aleksander Zgrzywa
Department of Information Systems, Faculty of Computer Science and Management
Wrocław University of Science and Technology
Wrocław
Poland

Andrzej Siemiński
Department of Information Systems, Faculty of Computer Science and Management
Wrocław University of Science and Technology
Wrocław
Poland

Kazimierz Choroś
Department of Information Systems, Faculty of Computer Science and Management
Wrocław University of Science and Technology
Wrocław
Poland

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-43981-5 ISBN 978-3-319-43982-2 (eBook)
DOI 10.1007/978-3-319-43982-2

Library of Congress Control Number: 2016947375

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

The International Conference "*Multimedia and Network Information Systems*"—MISSI 2016 organized by the Wrocław University of Science and Technology in Poland is an international scientific biennial conference for research on the field of multimedia and different aspects of Internet systems. This volume contains the proceedings of its 10th jubilee edition. The aim of the conference is to provide an internationally respected forum for research work on these areas. This conference covers their theoretical or practical aspects.

The MISSI Conferences are already a well-established event. During the span of the 20 years when the conference has been held we have the opportunity to see how the research area evolves, how new topics raised and the once formidable problems were solved. We have the privilege to host participants mostly from European countries but also from Asia, Africa, and America. The jubilee edition is no exception to the rule.

We have received 64 papers. All of them were reviewed by at least two members of the International Reviewer Board. The board has selected 36 papers with the highest quality for oral presentation and publication in the proceedings. All of them are original and were not published anywhere else.

This high turnout is probably due not only to the conference reputation but also due to the area it covers. The MISSI Conference is a specialized conference but the interest in multimedia and network systems does not vain. It grows with every year. Looking back over the previous editions of the conference we see that many of the subjects then fiercely discussed such as handwriting recognition or others are now embedded in everyday products.

The papers included in the proceedings cover among others the following topics: lip speech identification, user recognition based on fingerprint analysis, playing field detection in sports video shots, machine translation into mobile augmented reality systems, fast transmission of 3D computer graphics in the Web, brain computer interfaces for game control, prediction of topic popularity on online social networks, evaluating the quality of Internet LTE connections, movie recommendation using collaborative filtering, parallel implementations of ant colonies,

mitigating the dangerous activities of Web robots, various aspects of natural language processing with an emphasis on multilingual data access and statistical machine translation, question answering, named entity recognition, prediction of topics popularity on online social networks, RDF event stream processing, personalization of learning process using behavioral measures.

Some papers have theoretical character while others describe innovative applications. During the presentations and following discussions the presenters had the opportunity to exchange ideas and it was fascinating to see how these two approaches could interact and inspire each other.

Our special thanks go to the General Chair, Steering Committee, Program Chairs, Special Session Chairs, Organizing Chair, and Publicity Chair for their work for the conference. Without their help it would be impossible to guarantee the high quality of the papers presented at the conference.

We would like to express our thanks to the keynote speakers: Prof. Grzegorz Dobrowolski (AGH University of Science and Technology, Poland), Prof. Janusz Kacprzyk (Polish Academy of Sciences, Poland), and Prof. Gottfried Vossen (University of Münster, Germany) for their topical and inspiring talks.

Finally, we thank all the authors for their valuable contributions and all other participants of the conference. The success of the conference would not be possible without their involvement.

We sincerely hope that we have achieved our goal of providing the research community the up-to-date account on the work on rapidly evolving field of multimedia and Internet data processing. We would be pleased if the Conference could stimulate and inspire further research work on these areas.

This would be the greatest award for our effort on organizing the MISSI 2016 Conference.

Wrocław, Poland
June 2016

Aleksander Zgrzywa
Kazimierz Choroś
Andrzej Siemiński

Program Committee

Witold Abramowicz, Poznań University of Economics, Poland
Costin Badica, University of Craiova, Software Engineering Department, Romania
Kazimierz Choroś, Wrocław University of Science and Technology, Poland
Rozenn Dahyot, Trinity College Dublin, Irland
Paul Davidsson, Malmö University, Sweden
Luminita Dumitriu, Dunarea de Jos University, Romania
Bogdan Gabryś, Bournemouth University, UK
Michal Haindl, the Institute of Information Theory and Automation of the CAS, Poland
Mohamed Hassoun, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliotheques (ENSSIB), Villeurbanne, France
Zbigniew Huzar, Wrocław University of Science and Technology, Poland
Czesław Jedrzejek, Poznan University of Technology, Poland
Mohammed Khamadja, SP_Lab, Electroniques Department, Mentouri University, Constantine, Algeria
Marek Kopel, Wrocław University of Science and Technology, Poland
Bożena Kostek, Gdansk University of Technology, Poland
Ondrej Krejcar, University of Hradec Kralovee, the Czech Republic
Elżbieta Kukla, Wrocław University of Science and Technology
Sylvie Lainé-Cruzel, ERSICOM—Université Jean Moulin Lyon 3, France
Steffen Lamparter, Siemens AG, Corporate Technology, Germany
Mykhaylo Lobur, Lviv Politechnic National University, Ukraine
Zygmunt Mazur, Wrocław University of Science and Technology, Poland
Klaus Meyer-Wegener, University of Erlangen and Nuremberg, Germany
Davide Moroni, ISTI—National Research Council of Italy (CNR), Italy
Tadeusz Morzy, Poznan University of Technology, Poland
Mieczysław Muraszkiewicz, Warsaw University of Technology, Poland
Katarzyna Musiał-Gabryś, Bournemouth University, UK

Ngoc Thanh Nguyen, Wrocław University of Science and Technology, Poland
Piotr Odya, Gdansk University of Technology, Poland
Tarkko Oksala, Aalto University, Finland
Issa Panahi, University of Texas at Dallas, USA
Maria Pietruszka, University of Technology in Lodz, Poland
Gianni Ramponi, University of Trieste, Italy
Dimitr Ruta, British Telecom, UK
Henryk Rybiński, Warsaw University of Technology, Poland
Andrzej Siemiński, Wrocław University of Science and Technology, Poland
Janusz Sobecki, Wrocław University of Science and Technology, Poland
Piotr Szczuko, Gdansk University of Technology, Poland
Grzegorz Szwoch, Gdansk University of Technology, Poland
Maria Trocan, Institut Supérieur d'Électronique de Paris, France
Zygmunt Vetulani, Adam Mickiewicz University in Poznań, Poland
Aleksander Zgrzywa, Wrocław University of Science and Technology, Poland

MISSI 2016 List of Reviewers

Costin Badica, University of Craiova, Romania
Rozenn Dahyot, Trinity College Dublin, Ireland
Paul Davidsson, Belkinge Institute of Technology, Sweden
Luminita Dumitriu, Dunarea de Jos University, Romania
Michal Haindl, Czech Academy of Sciences, Prague, Czech Republic
Zbigniew Huzar, Wrocław University of Science and Technology, Poland
Czesław Jędrzejek, Poznań University of Economics, Poland
Marek Kopel, Wrocław University of Science and Technology, Poland
Bożena Kostek, Gdańsk University of Technology, Poland
Ondrej Krejcar, University of Hradec Králové, Czech Republic
Elľbieta Kukla, Wrocław University of Science and Technology, Poland
Klaus Meyer-Wegener, Friedrich Alexander University, Erlangen-Nuremberg, Germany
Davide Moroni, Institute of Information Science and Technologies ISTI-CNR, Italy
Tadeusz Morzy, Poznań University of Economics, Poland
Mieczysław Muraszkiewicz, Wrocław University of Science and Technology, Poland
Piotr Odya, Gdansk University of Technology, Poland
Tarkko Oksala, Helsinki University of Technology, Finland
Maria Pietruszka, Technical University of Lodz, Poland
Gianni Ramponi, Department of Industrial Engineering and Information Technology, University of Trieste, Italy
Janusz Sobecki, Wrocław University of Science and Technology, Poland

Piotr Szczuko, Gdansk University of Technology, Poland
Grzegorz Szwoch, Gdansk University of Technology, Poland
Ngoc Thanh Nguyen, Wrocław University of Science and Technology, Poland
Maria Trocan, Institut Supérieur d'Électronique de Paris (ISEP), France

Additional Reviewer

Andrzej Sikorski, Poznań University of Economics, Poland

Contents

Part I
Images and Videos Virtual and Augmented Reality

Building Knowledge for the Purpose of Lip Speech Identification

Andrzej Czyżewski, Bożena Kostek, Marcin Szykulski and Tomasz E. Ciszewski

Abstract Consecutive stages of building knowledge for automatic lip speech identification are shown in this study. The main objective is to prepare audio-visual material for phonetic analysis and transcription. First, approximately 260 sentences of natural English were prepared taking into account the frequencies of occurrence of all English phonemes. Five native speakers from different countries read the selected sentences in front of three cameras. Video signals, synchronized with audio, were registered and then analyzed. Encountered problems related to video registration and results achieved are discussed.

Keywords Audio-visual speech recognition · AVSR · Thermovision · Stereo-vision · Time-of-flight · Phonetic transcription

1 Introduction

Although in the recent years a huge progress has been made in the automatic speech recognition (ASR) [1, 2] there is still a place for further improvement, especially in cases when speech is produced casually or it is acquired in the presence of acoustical noise [3]. Moreover, as stated by Li et al. [4] advancement in that area is needed under real-world adverse conditions, especially as the range of ASR-based

A. Czyżewski · B. Kostek · M. Szykulski (✉) · T.E. Ciszewski
Faculty of Electronics, Telecommunications and Informatics, Multimedia Systems, Department and Audio Acoustics Laboratory, Gdansk University of Technology, Narutowicza 11/12 80-233, Gdańsk, Poland
e-mail: marszyk@sound.eti.pg.gda.pl

A. Czyżewski
e-mail: andcz@sound.eti.pg.gda.pl

B. Kostek
e-mail: bozenka@sound.eti.pg.gda.pl

T.E. Ciszewski
e-mail: tomasz.e.ciszewski@gmail.com

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*, Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_1

applications grew significantly, covering robotics, human-computer interfaces, car voice navigation [5], and automatic transcription for people hearing impairments.

The main aim of the research described in this paper is to build background knowledge on joint audio-video speech recordings serving for the purpose of developing methods for bi-modal (audio-video) lip speech detection and transcription. Several speakers read a list of sentences which were recorded for this study by three cameras: stereoscopic (ST), thermovision (TH) and Time-of-Flight (ToF). All video recordings were synchronized with the audio signal recorded via two microphones. However, some related difficulties were encountered. To build the knowledge, first all the recordings were processed and indexed and then a phonetic specialist's task was to transcribe and analyze the words uttered by the volunteer speakers. Further, the recordings were parametrized. In Sect. 2 the assumptions and the detailed experimental setup are presented. Section 3 contains the examples of analyses along with the discussion on the performed experiments. This is followed by concluding remarks which point out the main difficulties encountered in the present study.

2 Speech Video- and Audio-Based Recordings

Several steps were devised in order to prepare audio-video material for further analysis. Thus, for the purpose of this analysis video- and audio recordings of natural English sentences were made. First, it was assumed that a sufficient language sample for training algorithms is approximately 250–300 sentences, whose average length is six lexical items [6]. All words from the first 1000 of Leech et al. [7] frequency list were used. This guarantees that the language sample includes most commonly used words. Some words from 2nd, 3rd and–sparingly–4th thousand sets were used in order to 'fine-tune' the source sound frequencies. The aim was to use each word only once. However, for obvious reasons, function words had to be used more than once, e.g. *the* (108 times), *a* (47 times). Some content words, e.g. *decision* (which was used twice), also had to be 'recycled' since their phonetic make-up was necessary, again, in order to fine-tune, or to approximate, the intended sound frequencies. Altogether, 168 words (both function and content) were used more than once. The target sentences had naturally varied prosody, which was achieved by using: a variable number of words in a sentence, words of different number of syllables, sentences of variable grammatical structure (statements (193), questions (16), negatives (15), imperatives (35) and exclamatory sentences (5)). Consequently, the gathered material includes approximately 260 sentences which were read by five native speakers originating from different countries.

In Fig. 1 the block diagram of the experimental setup, and in Fig. 2 the layout of the cameras employed in the recordings are shown. In Table 1 the specification of the three cameras employed in the experiments are presented. As shown in Fig. 1, the stereoscopic camera enables to register both audio and video, simultaneously. Moreover, such a type of camera sends synchronization timecode via the HDMI

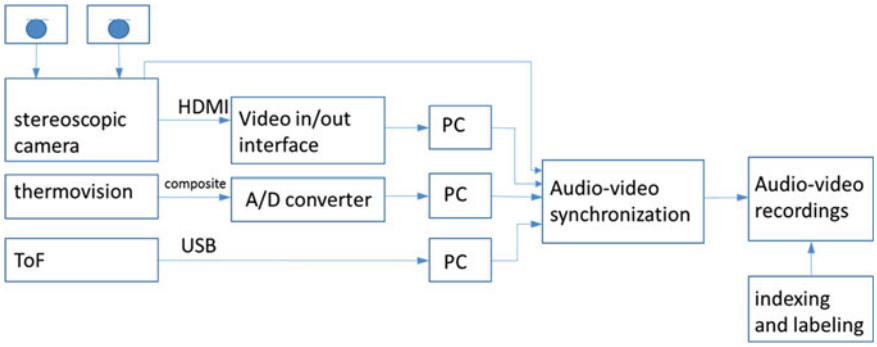


Fig. 1 Block diagram of the experimental setup

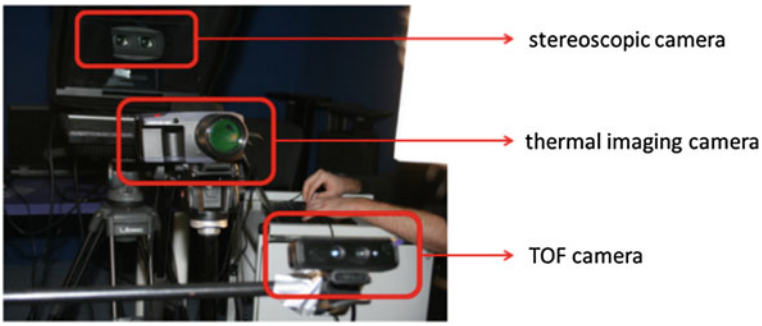


Fig. 2 Layout of three cameras employed in video recordings

Table 1 Cameras specification

Camera type	Spatial resolution	Time resolution (FPS)
Stereoscopic (ST)	1280 × 720	50
Thermovision (TH)	320 × 240	30
ToF	320 × 240	60

port. On the other hand, the thermovision camera needed an additional Analog-to-Digital (A/D) converter. As explained further, this also caused synchronization problems and therefore the recordings obtained in this way had to be manually edited and indexed using, among others, the VirtualDub program.

The characteristics of recorded modalities were as follows: **Audio**—reference soundtracks (separate channels L and R); **Stereo Vision**—video track recorded with

the stereoscopic camera (separate L and R channels); **Thermovision**—video track recorded with infrared camera; **ToF**—Time of Flight camera.

It turned out that the processing of the recordings obtained from the thermovision camera using H.264 codec was ineffective. The video material obtained in this way had a variable framerate. This problem has forced the authors to perform a separate synchronization for each recorded word. For this purpose the VirtualDub program was employed, which, however, resulted in a separate file for each indexed word. Therefore, why realization of further recordings enforced the authors to use the wmv2 source encoder. In this way the process was reduced to a few steps. The first step was to change the codec to H.264 without compromising on the quality with the program ffmepg. Then, it was possible to adjust the reference audio track to the entire recording in the program VirtualDub. This enabled the indexing of the whole file, without having to synchronize individual words. The recordings from ToF cameras were the most problematic in the context of establishing a uniform framerate. This means that it was necessary to use an identical approach to thermal imaging and H.264 codec. Moreover, at this stage the process was not automated. Overall, the process of creating and indexing a sample multimodal database record has proved to be long and tedious. The reasons were: the time-consuming processes associated with the way the acquisition of recordings for each modality, and more precisely the lack of synchronization throughout the entire recording. In the case of infrared camera (H.264 encoding) and ToF camera, the preparation of a single set of files took about 10–11 h. Moreover, as for each modality (except ToF) three series of recordings were made, which substantially increased the total effort. It was also noted that the indexing process based solely on the audio input resulted in the loss of audio information. Such case is illustrated in Fig. 3 where video frames precede the corresponding audio signal, which fact should be taken into account in the indexing process.

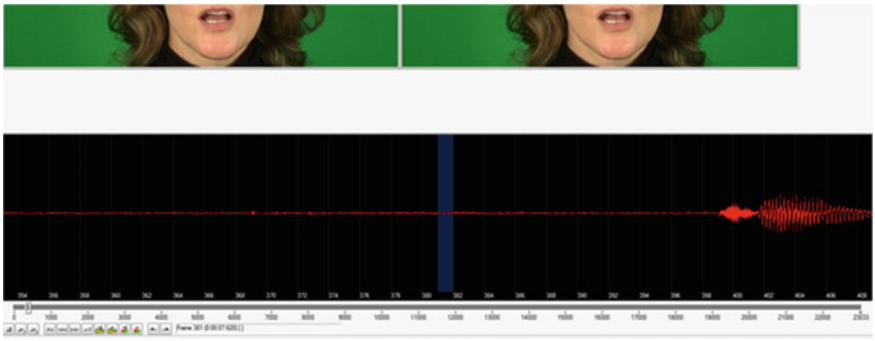


Fig. 3 Illustration of the need for video frames preceding the corresponding audio signal

3 Experiments and Discussion

Several assumptions were made about the visibility of particular phonemes in the video recordings at the early stage of the experiments [6]. Prior to the recordings, a phonetic expert's task was to formulate some rules with regard to recognizing phonemes visible in video frames coming from different cameras. However, it was observed that casual pronunciation of particular words may make these rules obsolete. In Fig. 4 video frames obtained from two speakers employing stereo vision (ST), thermovision (TH) and ToF cameras, as well as the corresponding audio signal along with the waterfall spectrogram are shown for the word: *bet*. In Fig. 5 similar analyses are presented for the word: *bit*. It may be observed that the word uttered by Speaker 1 is more careful and exact, thus being close to what had been assumed before the recordings took place.

Also, some other problems were encountered pertaining to the visibility of articulatory speech organs. For example, the inevitable movements of the speaker's head changing the exposition angle in relation to the centrally positioned camera in a standard PC or laptop are visible in Fig. 5 (Speaker 4, word: *bit*).

Moreover, speaker individual facial expressiveness (or the lack of it) and the speaking style may affect the transcription of the utterances. As may be observed, a problem with interpreting the thermal imaging recording seems also important.

In the video frame with a regular, stereoscopic camera (ST) both vowels /i/ and /e/ are more visible than it was expected at an earlier stage. In particular, this happens if the lips for these two vowels are juxtaposed. This is also noticeable in the image of the infrared camera and ToF, although the ToF image is in general not very clear.

For the vowel /e/ in the word the word *bet* (SPEAKER 1 and SPEAKER 4) the following observation have been made: ST—clearly visible (better than expected), TH—confirmed (poor visibility, the lip setting suggests their rounding, which is not present in the articulation of this vowel), ToF—confirmed as above. In turn, for the vowel /i/ in the word *bit* (SPEAKER 1 and SPEAKER 4) it was observed that: ST—clearly visible (better than expected), TH—confirmed as above.

Figures 6 and 7 present spectrograms corresponding to the words *bet* and *bit* aligned with their phonetic transcription for Speakers 1 and 4.

Further analysis consisted in parametrization of the indexed audio and video recordings, i.e. performing mel-cepstral analysis for the audio signal and AAM (Active Appearance Models) for the video recordings [8–10]. Several (39) MCFC coefficients were calculated, of which only the first one is presented for both speakers (1 and 4) and for both words (*bet* and *bit*), see Fig. 8.

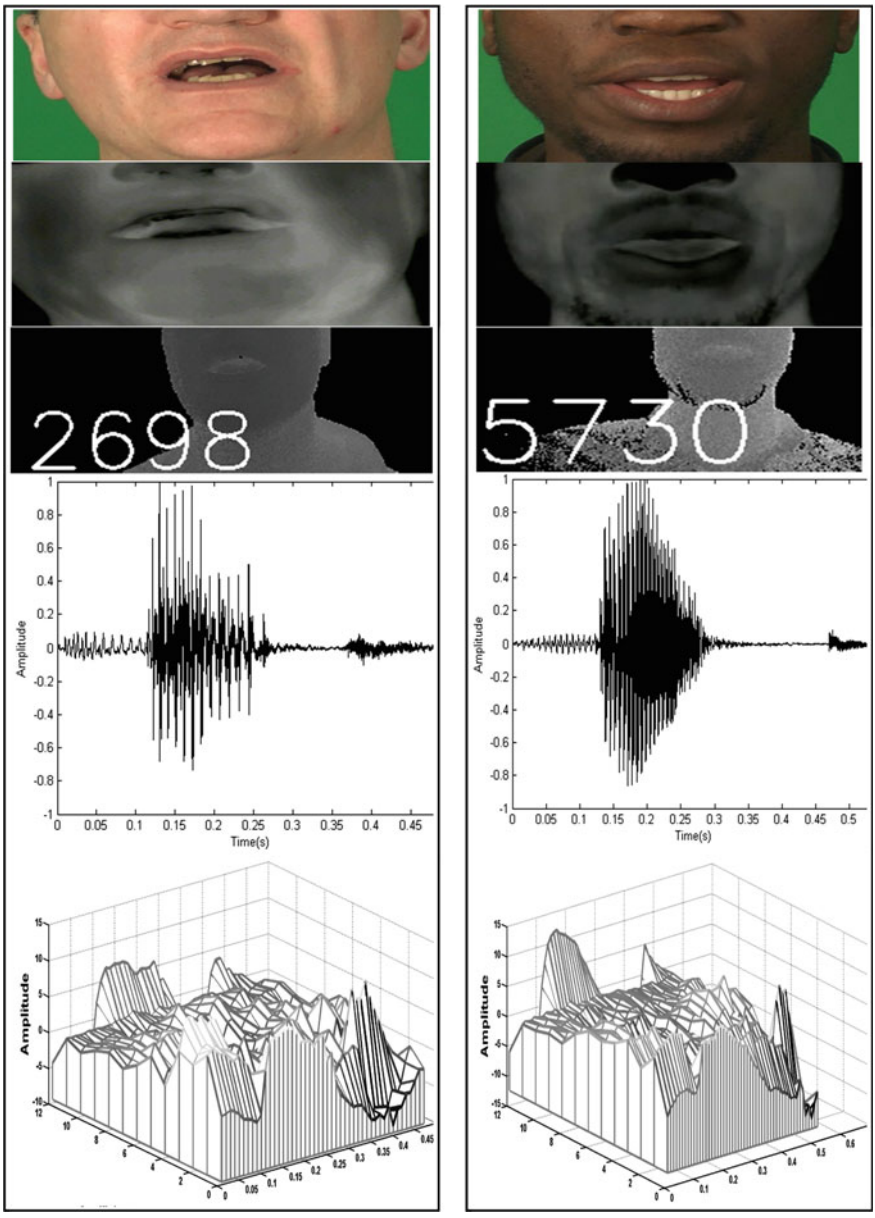


Fig. 4 Speaker 1 (*left side*), Speaker 4 (*right side*), word: *bet*, from *top* to *bottom*: a video frame from stereo vision, thermovision, ToF (frames no. 2698 and 5730) cameras, corresponding audio signal and waterfall spectrogram

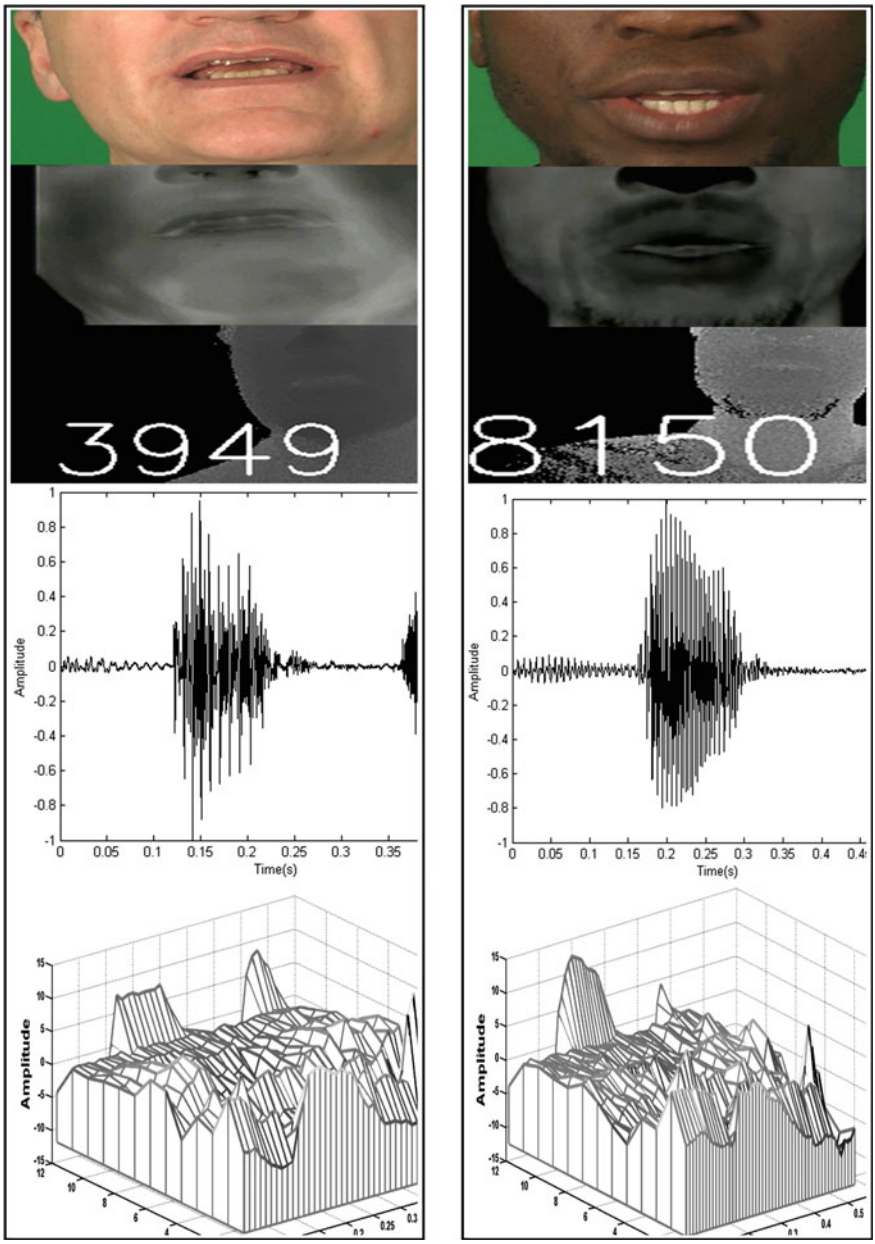


Fig. 5 Speaker 1 (*left side*), Speaker 4 (*right side*), word: *bit*, from *top* to *bottom*: a video frame from stereovision, thermovision and ToF (frames no. 3949 and 8150) cameras, corresponding to the audio signal and the waterfall spectrogram

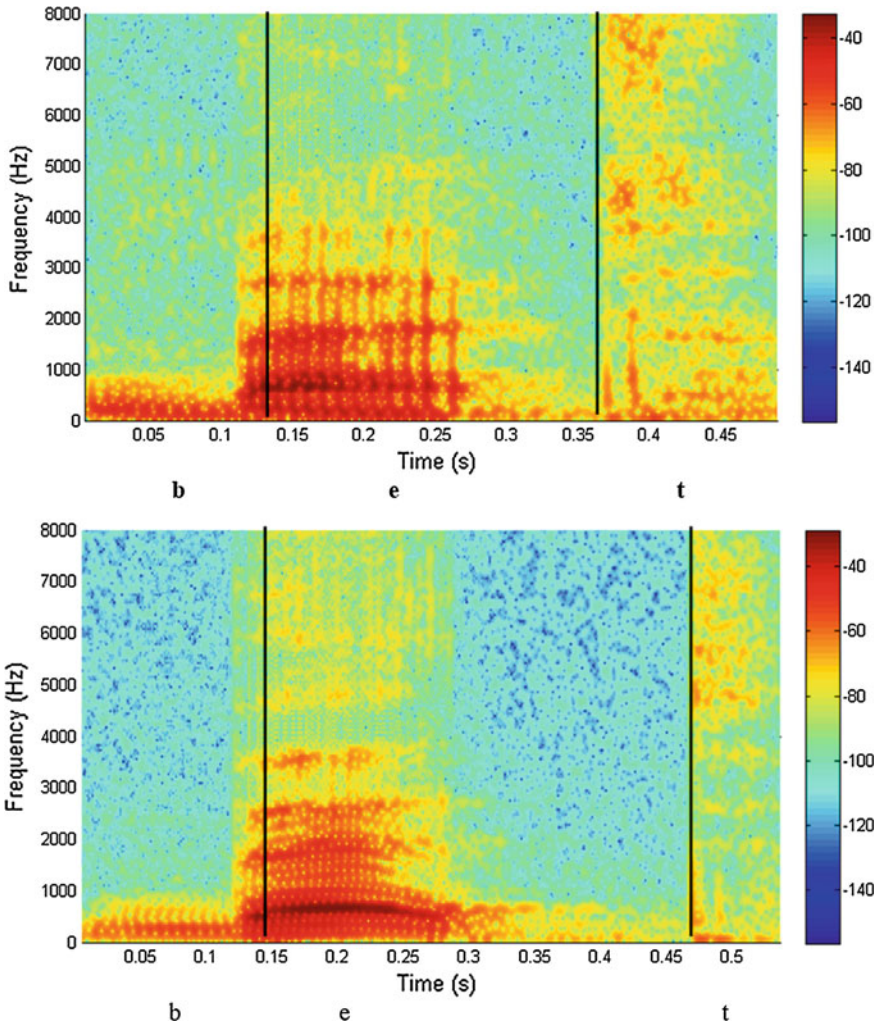


Fig. 6 Spectrogram of the word *bet* aligned with phonetic transcription, Speaker 1 (*top*), Speaker 4 (*bottom*)

As mentioned before, the speaker's 4 articulation manner was much more casual, thus the analyses performed differ between these two speakers, as well as among other recorded volunteers. The parametrization of video recordings was presented in earlier papers by the Multimedia Systems Department team of Gdansk University

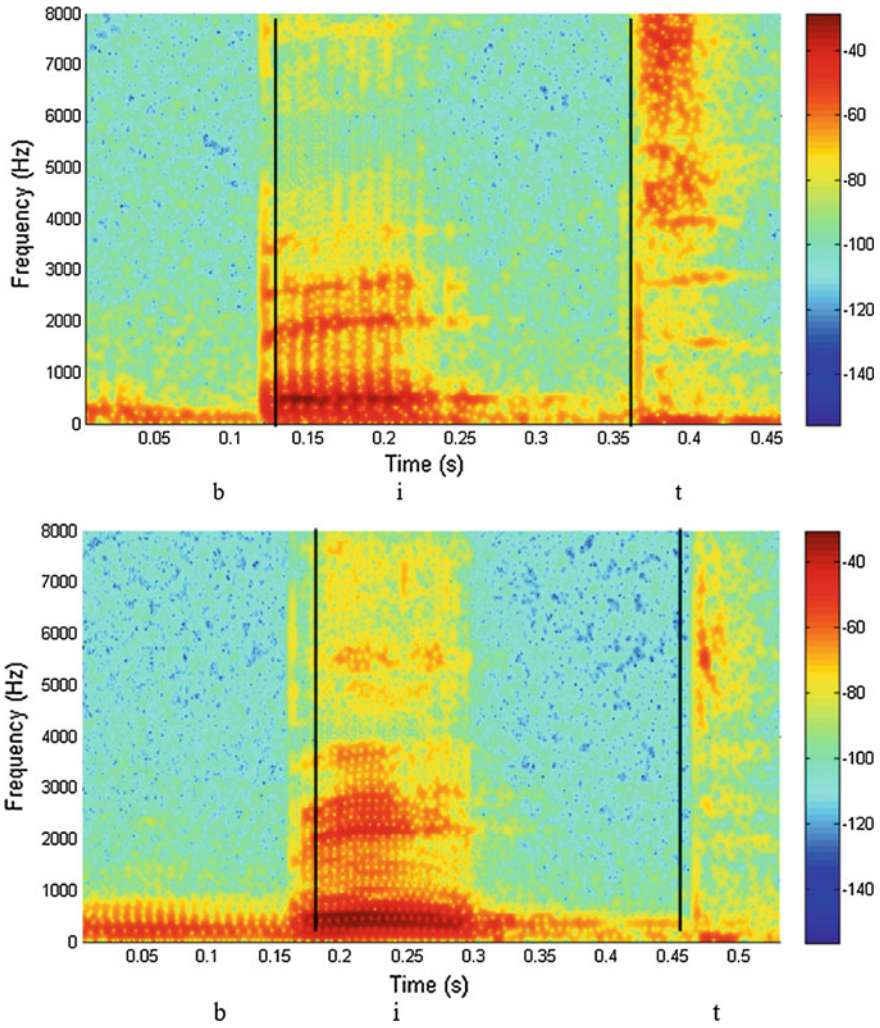


Fig. 7 Spectrogram of the word *bit* aligned with phonetic transcription, Speaker 1 (*top*), Speaker 4 (*bottom*)

of Technology, and will not be recalled here [8–10], as all stages consisting in face and lip identification, as well as the static (frame-based) and the dynamic feature vectors derivation, were discussed thoroughly elsewhere.

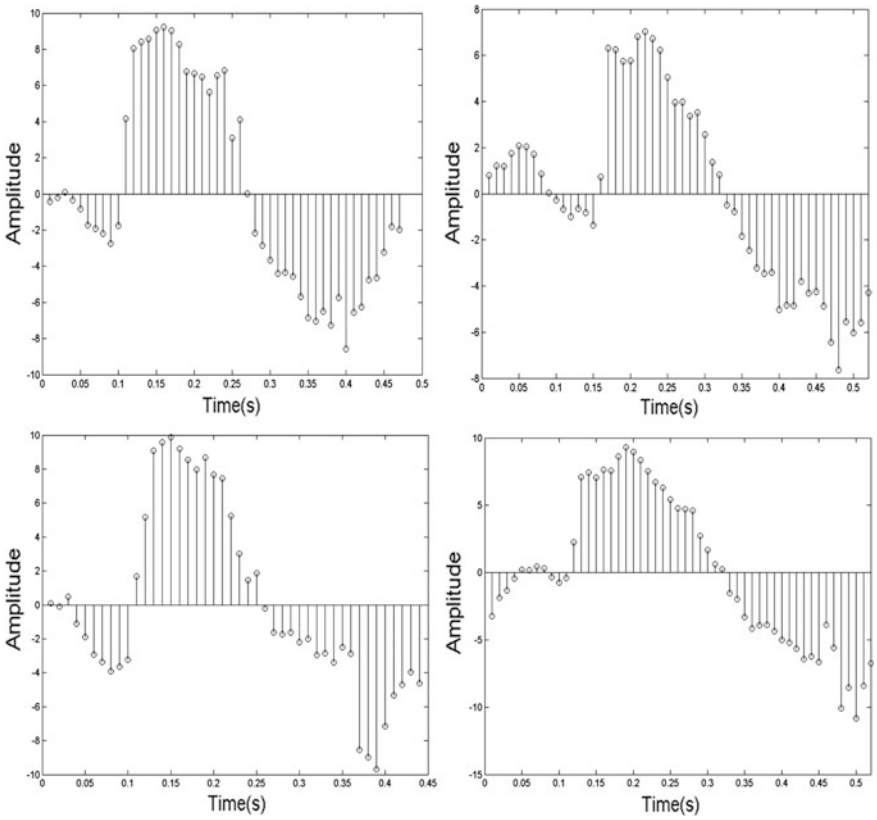


Fig. 8 MFCC (first) for Speaker 1 (*left side*), Speaker 4 (*right side*) of the word *bet* (*top*) and *bit* (*bottom*)

4 Conclusions

An attempt in multimodal registering of English speech was made. Several problems were encountered during the performed experiments with the simultaneous video- and audio-recordings. The most important problem at this stage of the study was the lack of technical means to ensure synchronicity between video frames and the corresponding acoustic signal in the case of thermovision and ToF cameras used in the recordings. Also, separating video frames adequately with the aligned audio signal has proved to be a difficult task, as it occurred that this process cannot easily be automated. The reason, among others, was that video frames preceded the corresponding words in the audio signal. This phenomenon was pointed out by Lavagetto [11], whose study demonstrated that acoustic and visual speech stimuli are asynchronous. This is easily explicable by the fact that visual articulators before producing speech have to position themselves correctly before and after the start

and end of an acoustic utterance. This phenomenon is called as the voice-onset-time (VOT) in the literature [12].

Some other problems occurred with the visual input recognition that may make the phonetic transcription difficult, i.e. single visible articulatory gesture spreading onto two or more phonemically distinct sound segments (e.g. pre-rounding/pre-spreading of the speaker's lips in anticipation of the following rounded/spread vowel, respectively) or identical articulatory gestures (lip shape, degree of opening), which accompany different types of sounds. This may serve as another proof that syntactic predictability can influence word recognition and may be less reliable, especially in the case of casual or conversational speech.

At this stage of the study, there are no conclusive findings concerning the integration of audio and video modalities in the context of bringing a significant improvement in automatic transcription of the collected material.

It was found however, that every modality provided some unique features that were absent in other streams. The high-resolution stereoscopic camera captured not only the movement of speech organs, but also the facial expressions of speakers, which differed depending on the sentence context (questions, negatives, imperatives, etc.). Thermovision data—despite lower resolution—revealed additional phonological information—e.g. identification of nasal consonants which was indicated by the hot air exhaled by the speaker's nostrils. In the case of nasal consonants, the mouth is occluded at some point by the lips or tongue and the airstream is expelled entirely through the nose, thus this effect is visible only in the thermal imaging camera. This feature may be important to distinguish between native and non-native speakers [13] or to detect consonants and vowels, nasal consonants and nasalized vowels, etc. [14]. The ToF camera, which captured the speaker's image from shoulders up, could be used in more casual recording conditions—to capture additional, non-facial articulatory gestures that are often associated with affected speech.

That is why the planned future research will aim at automatic indexing and searching for optimum parametrization and fusion of modalities, as well as checking the robustness of lip speech decoding in noisy conditions. It should also be remembered that acoustic and visual modalities can be combined either at the feature or the decision level [15]. These two integration models, referring to early and late integration of bimodal information may strongly influence the classification process as they result either in parallel processing with separate decisions or in a so-called coactive processing, both strategies having advantages and disadvantages.

Acknowledgments Research sponsored by the Polish National Science Centre, Dec. No. 2015/17/B/ST6/01874.

References

1. Potamianos, G.: Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE **91**(9), 1306–1326 (2003) doi:10.1109/JPROC.2003.817150
2. Zhou, Z., Zhao, G., Hong, X., Pietikäinen, M.: A review of recent advances in visual speech decoding. Image Visual Comput. **32**(9), 590–605 (2014)
3. Zeiler, S., Nicheli, R., Ma, N., Brown, G.J., Kolossa, D.: Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis. ICASSP **2016**, 2797–2801 (2016)
4. Li, J., Deng, L., Gong, Y., Haeb-Umbach, R.: An overview of noise-robust automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. **22**(4), 745–777 (2014)
5. Biswas, A., Sahu, P.K., Chandra, M.: Multiple camera in car audio–visual speech recognition using phonetic and visemic information. Comput. Electr. Eng. **47**, 35–50 (2015)
6. Czyżewski, A., Kostek, B., Ciszewski, T., Majewicz, D.: Language material for English audiovisual speech recognition system development. In: J. Acoust. Soc. Am. **134/5**, 4069 (2013) (abstr.) and Proc. Meet. Acoust., **1**(20), 1–7, San Francisco, USA, 2.12.2013–6.12.2013
7. Leech, G., Rayson, P., Wilson, A.: Word Frequencies in Written and Spoken English: Based on the British National Corpus. Longman, London (2001)
8. Kunka, B., Kupryjanow, A., Dalka, P., Bratoszewski, P., Szczodrak, M., Spaleniak, P., Szykulski, M., Czyżewski, A.: Multimodal English corpus for automatic speech recognition. In: IEEE Signal Processing—Algorithms, Archit. Architectures, Arrangements, and Applications Conference. Proceedings, Poznań, Poland, 26–28, (2013)
9. Dalka, P., Bratoszewski, P., Czyżewski, A.: Visual lip contour detection for the purpose of speech recognition. In: International Conference on Signals and Electronic Systems, Poznań (2014)
10. Bratoszewski, P., Czyżewski, A.: Face profile view retrieval using time of flight camera image analysis. In: Pattern Recognition and Machine Intelligence 2015, vol. 9124, pp. 159–168, Warsaw (2015)
11. Lavagetto, F.: Converting speech into lip movements: a multimedia telephone for hard hearing people. IEEE Trans. Rehab. Eng. **3**(1), 90–102 (1995)
12. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia **2**(3), 141–151 (2000)
13. Harnsberger, J.D.: On the relationship between identification and discrimination of non-native nasal consonants. J. Acoust. Soc. Am. **110**(1), 489–503 (2001)
14. Marques, L.F.: Variation in nasality between nasal and nasalized vowels in Brazilian portuguese: a pilot study. J. Inicio. **9**(1) (2014). http://ojs.gc.cuny.edu/index.php/lljournal/article/view/1446/1546. Accessed May 2016
15. Lucey, S., Chen, T., Sridharan, S., Chandran, V.: Integration strategies for audio-visual speech processing: applied to text dependent speaker recognition. IEEE Trans. Multimedia **7**(3), 495–506 (2005). doi:10.1109/TMM.2005.846777

PNG as Fast Transmission Format for 3D Computer Graphics in the Web

Daniel Dworak and Maria Pietruszka

Abstract This paper focuses on manners of filling the gaps in existing standards that are used in tridimensional Web technologies. We proposed the way of encoding huge 3D data sets in lossless PNG format and use of programmable rendering pipeline to decoding PNG file. It allows to reduce significantly the file with 3D data, time of transmission via Web and time needed to decode the file.

Keywords 3D geometry in the Web · 3D data transmission format · WebGL rendering pipeline · GPGPU

1 Introduction

Nowadays, every personal computer, notebook or even mobile device has its own graphics processing unit (GPU) and installed newest Web browser, which has an access to GPU via shaders language. It allows to perform complex calculations using GPU to render tridimensional graphics data in real time. GPU shaders can be also used for non-graphics data-parallel computing, for example image, audio and video processing [3], which is called GPGPU—General-Purpose Computing on Graphics Processing Units.

The main problem for 3D graphics in the Web are huge 3D data sets, which contain many models of buildings, trees, etc. Transmission of such data via network requires wideband Internet connection and transmission formats. There are transmission formats for Audio (MP3), Video (H.264), and images (JPEG, PNG) but no

D. Dworak (✉) · M. Pietruszka (✉)
Institute of Information Technology, Lodz University of Technology, Lodz, Poland
e-mail: 150859@edu.p.lodz.pl
URL: http://www.p.lodz.pl

M. Pietruszka
e-mail: maria.pietruszka@p.lodz.pl

D. Dworak
Center for Media and Interactivity, Justus Liebig University Giessen, Giessen, Germany

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_2

transmission format exists for 3D computer graphics yet. Declarative-3D formats, such as X3D and Collada, specify 3D scene data in a well-structured, human-readeable format. However, a text-based meshs data description gives huge files, even for small models that consists of a few thousand polygons. WRL was declarative-3D format of VRML (Virtual Reality Markup Language), first 3D graphics technology for the Web. The X3D technology, as a successor to VRML, accepts description of 3D scene in old WRL format and in new XML format, but in both cases, the additional plug-ins are necessary. Collada is also declarative-3D asset interchange format (DAE), created by the Khronos Group.

Currently, there are several solutions to convert those formats to modern Web standards without plug-ins. X3DOM is an open-source and runtime framework for 3D graphics for the Web, which allows to include X3D elements as a part of any HTML5 DOM tree [1]. XML3D offers another approach for interactive 3D graphics (encoded in XML), that is not based on any prior standard [11]. More general solution is the Khronos Group's glTF format (WebGL Transmission Format), through which it is possible to use Collada format prepared for current WebGL technology (Fig. 1) [7, 12]. A glTF asset is represented by JSON file (for node hierarchy, materials, cameras, as well as descriptor information for meshes, shaders, animations and other construct), binary files (for geometry and animations, and other buffer-based data), JPEG and PNG image files for textures, and GLSL text file for shader source code.

glTF like X3DOM uses text encoding for structured declarative data (such as transformation groups, materials, cameras, and descriptor information about meshes) and binary encoding for unstructured data for geometry (the vertex attributes of the mesh). Unstructured data usually constitutes more than 95 percent of a file, what affects the data rate. Limper et al. developed two binary 3D mesh formats for declarative 3D approaches: sequential image geometry (SIG) using PNG images as a regular mesh container, and progressive binary geometry (PBG) using a direct binary encoding of mesh data [8]. Our proposition was PNG as transmission format for irregular meshes (Fig. 2) [2]. The research has shown that the 3D data in PNG files reduces its size and time needed to transfer them via Internet.

This paper focuses on the 3D data encoding in PNG file and on data decoding simultaneously. For this purpose, it has been proposed to use a GPU rendering pipeline supported by WebGL technology [4, 9, 10]. Concerning a decoding of

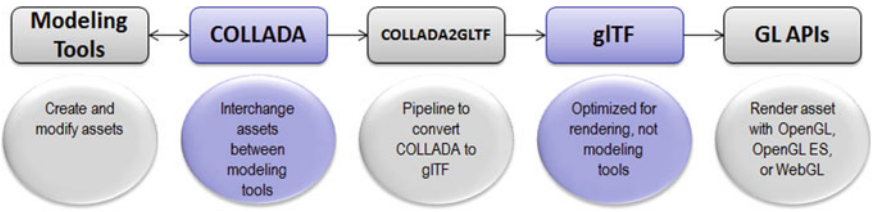


Fig. 1 Khronos Groups glTF as transmission format for WebGL aplications [5]

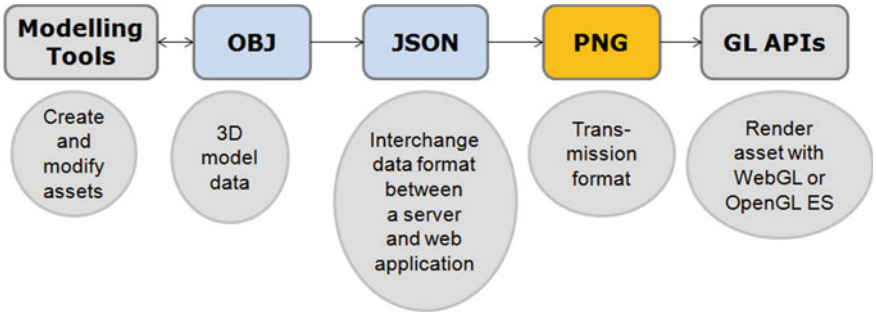


Fig. 2 PNG as transmission format for 3D Web applications

large two-dimensional sets, the most important thing is the accessible, high clocked, large memory of GPUs (even up to 8GB for single GPU), which can be used by Web browser easily and fast. The architecture of graphics card allows to perform a lot of operations on the whole data set in one unit of time, which is called calculations parallelism, however it is hard to control the order of tasks, as well. Moreover, there are no dependencies between data elements, therefore it is impossible to addict one value to another, what also makes it hard (or even sometimes not possible) to redesign traditional algorithms.

2 Encoding 3D Data to PNG File

Basically, there are five types of data to define 3D model: vertices (space coordinates x, y and z), faces (set of vertices indices which creates single polygon), normals (vectors of surfaces orientation for lighting calculations), textures coordinates (called UV's), material properties (like color coefficients and images for texture). We have been decided to use JSON (JavaScript Object Notation) format for storing 3D data. This format is an open standard with user-friendly notation (e.g. readable text) and is used to transmit data from server to client easily. The JSON file has a structure like follows [6]:

Listing 1. 3D data for graphics scene:

1. **List of materials:** "*materials*" : [{"*DbgIndex*" : 0, "*DbgName*" : "*sample*", "*colorAmbient*" : [0.925, 0.807, 0.164], "*colorDiffuse*" : [0.925, 0.807, 0.164], "*colorSpecular*" : [0.0, 0.0, 0.0], "*transparency*" : 0.0, "*mapAmbient*" : "1.*png*", "*mapDiffuse*" : "1.*png*", }]
2. **Vertices**: "vertices": [–5.0, 0.0, 5.0, 5.0, 0.0, 5.0, –5.0, 0.0, –5.0, 5.0, 0.0, –5.0, –5.0, 5.0, . . . , –5.0],
3. **Normals coordinates**: "normals": [0.0, –1.0, 0.0, 0.0, –1.0, 0.0, 0.0, 1.0, 0.0, 0.0, 1.0, –1.0, . . . , 0.0],

4. **Texture's coordinates**: "uvs": [0.0, 0.0, 1.0, 0.0, 0.0, 1.0, 1.0, 1.0, 0.0, 0.0, 1.0, 0.0, 0.0, ..., 1.0],
5. **Faces**: "faces": [42, 0, 2, 3, 0, 9, 11, 10, 0, 0, 0, 42, 3, 1, 0, 0, ..., 11, 11, 11]

The conception of conversion JSON to PNG files is based on saving any 3D data to RGB channels of 2D image (Fig. 3). Then, every value (vertex, normal, UVs) is splitted into integer and fractional parts and stored in R, G, B channels. The Alpha channel (A) identifies kind of data: vertex ($A = 128$), faces ($A = 100$), normal ($A = 50$), texture coordinate ($A = 25$), material ($A = 255$).

To store a single vertex coordinate it is necessary to normalize every vertex value to interval from 0 to 255, then we perform calculations as follows:

$$R = \lfloor x_{position} \rfloor \tag{1}$$

$$t = \frac{(x_{position} - R) * 10000}{256}, G = \lfloor t \rfloor \tag{2}$$

$$B = t - G \tag{3}$$

$$A = 128. \tag{4}$$

where $\lfloor a \rfloor$ is a floor of a.

Information about faces are slightly unusually encoded. The face flag and indices are encoded one by one for vertices, material, texture UV's, normals and color. The flag is encoded as follows [6]:

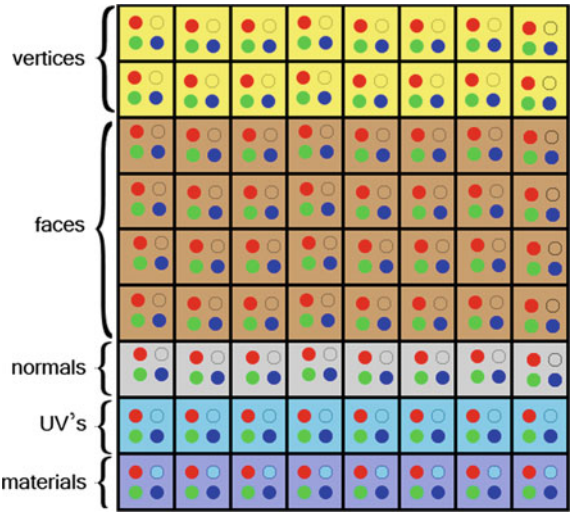


Fig. 3 Structure of PNG file with geometry—vertices (*yellow background*), faces (*light brown*), normals (*gray*), UVs (*light blue*) and materials (*light violet*)

1. bit—face type (0 if quad, 1 if triangle)
2. bit—1 if face has material, 0 otherwise
3. bit—1 if face has UV's, 0 otherwise
4. bit—1 if vertices have UV's, 0 otherwise
5. bit—1 if face has normals, 0 otherwise
6. bit—1 if vertices have normals, 0 otherwise
7. bit—1 if face has colors, 0 otherwise
8. bit—1 if vertices have colors, 0 otherwise

Every face value (flag and indices of vertices, materials, normals and UV's) is an integer number and requires only two channels—R and G:

$$R = \lfloor \frac{face_{value}}{256} \rfloor \tag{5}$$

$$G = \frac{face_{value}}{256} - R,\ B = 0 \tag{6}$$

$$A = 100 \tag{7}$$

where $face_{value}$ is every integer number for faces set.

Storing of normals is similar to faces, but we also store a sign of the value in blue channel.

$$t = \frac{normal_{value} * 10000}{256}, R = \lfloor t \rfloor \tag{8}$$

$$G = t - R \tag{9}$$

$$A = 50 \tag{10}$$

If $normal_{value} \geq 0$, then $B = 255$, otherwise $B = 0$.

And finally, to store a UV's texture value is similar to a normal value, but UV's are not signed.

$$t = \frac{uv_{value} * 10000}{256}, R = \lfloor t \rfloor \tag{11}$$

$$G = t - R,\ B = 0 \tag{12}$$

$$A = 25 \tag{13}$$

At the end of PNG file, informations about materials are saved. There has been proposed JSON based sequence of properties like: materials' name, colors, textures' names or transparency. A textual information (for example the name of a texture map) is stored in ASCII code—one character in single pixel's channel but material's colors—as RGB pixels.

3 Decoding 2D Data Sets to 3D Geometry

The 3D data saved in PNG file are decoded according to equations as follows:

$$Vertex_{value} = R + \frac{G * 256 + B}{10000}, \quad Face_{value} = R + G * 256,$$
$$Normal_{value} = \frac{(R + G * 256)}{10000}, \quad UV_{value} = \frac{R + G * 256}{10000}. \tag{14}$$

Next, they need to be saved in Vertex Buffer, from where they are transferred to rendering pipeline. The current paradigm for the Web is programmable rendering pipeline (Fig. 4). However, there is no ability to use the latest technologies which are offered by graphics cards (using by 20 % of users), because the Web application has to be run on the great number of computers. Nowadays it is possible to program two shaders. Vertex Shader processes a single vertex, while Fragment Shader a single fragment, which corresponds with a pixel of the output file. Data for Vertex Shader comes from Vertex Buffer, to which they are saved by 3D application. The output data of Fragment Shader's computations are sent to Framebuffer—the place from wheres are taken to display. In WebGL technology exists Frame Buffer Object (FBO), which allows to create a user-defined invisible Framebuffer. It allows to save floating point numbers (processed by Fragment Shader) in such FBO, which contain decoded 3D data (Eq. 14).

In this article, the main goal of operations performed in rendering pipeline is not a displaying the output PNG picture, but decoding data about vertices, faces, normals and UVs saved there (Fig. 5). It can be achieved by mapping PNG picture on a square. In Fragment Shader, the calculations accoring to equations (14) are performed and the results are saved in textures texels, which are the floating point numbers. The final output of rendering process is directed to invisible Framebuffer, instead of displaying it on the screen.

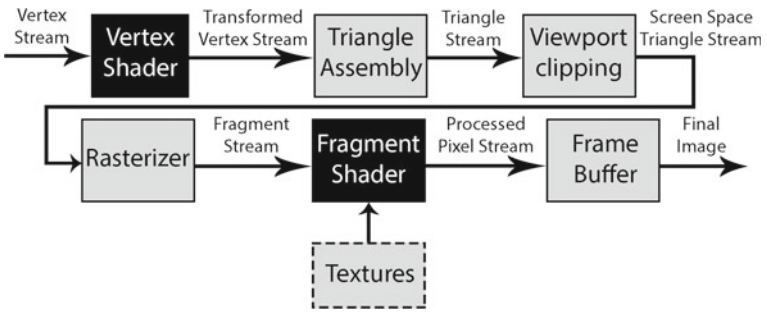


Fig. 4 Programmable rendering pipeline for the web

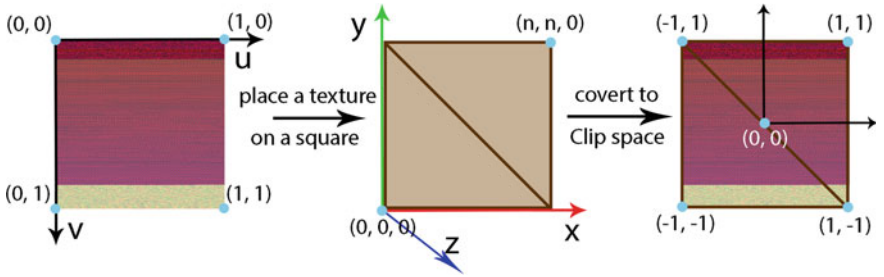


Fig. 5 Programmable rendering pipeline for decode 3D data from PNG file; n—dimension of PNG image

This algorithm defines stages of decoding 3D graphic's data stored in PNG file:

(1) Load PNG file.
(2) Define parameters of a texture created from a picture saved in PNG file.
(3) Save data needed to render a square in Vertex Buffer.
(4) Draw the square, on which an image from PNG file is mapped.
(5) Save the results from invisible Framebuffer into arrays of geometry's data.

Only the fourth stage is run on GPU, the other stages are run on CPU.

We have implemented this algorithm using WebGL, that brings hardware-accelerated cross platform and hardware independent 3D graphics to the browser without installing additional software (e.g. plugin). For this reasons and compatibility with HTML5 language it is the most popular nowadays. To perform some actions with WebGL, it requires to create the canvas element which is 2D drawing context API.

Re 1. After loading the image to CPU, special buffer named Frame Buffer Object (FBO) and a context (on the canvas element) need to be initialized.

Re 2. The parameters for wrapping the texture are set to *GL_CLAMP_TO_EDGE* and *GL_NEAREST* for mapping; the rest are defaults.

Re 3. The square with dimensions of the PNG image is created from two triangles (Fig. 5). The origin of the image is located in the upper left corner, while the origin of scene's coordinates is located in the bottom left corner, therefore to define the right texture coordinates it is necessary to mirror vertical axis. Then, a vertex located in $(0, 0, 0)$ corresponds with a UV $(0, 1)$, whereas a vertex $(n, n, 0)$ corresponds with UV $(1, 0)$.

Re 4. In WebGL, it is possible to use *drawArray* function, which renders triangle primitives from the array to binded context < *canvas* > element. Then, the programmable rendering pipeline is launched, for which it is needed to create Vertex Shader and Fragment Shader (Fig. 5). Moreover, it is possible to build a lookup table to ensure high efficiency while determining position of each pixel.

In WebGL technology, the Shaders are programmed in the OpenGL Shading Language (GLSL) [9]. The Vertex Shader, in case of our researches, computes

textures coordinates (Listing 2). The input data are position and UVs coordinates for the current vertex and dimension of PNG image (*u_resolution*). The space position coordinates (eg. x, y and z) needs to be transposed to Clip space coordinates (interval [1; 1]) due to requirements of *gl_Position* variable. At the end, the *gl_Position* and *v_texCoord* are redirected to Fragment Shader. Input data for Fragment Shader is also the texture image (*u_image*). Next, according to alpha value (converted to interval [0.0, 1.0]) and Eqs. (14) a data about 3D scene are calculated using Fragment Shader (Listing 3).

Listing 2. An example of Vertex Shader that handles the texture and passes it to Fragment Shader

```
<script id="vertexShader"type="x-shader/x-vertex">
attribute vec2 a_position; //position of the current vertex/pointer
attribute vec2 a_texCoord; //texture's UV's
uniform vec2 u_resolution; //texture's resolution
varying vec2 v_texCoord; //coordinates for Fragment Shader

void main() {
  vec2 zeroToOne = a_position / u_resolution; //convert pixels'
                                                rectangle to [0.0, 1.0]
  vec2 zeroToTwo = zeroToOne * 2.0; // convert from 0->1 to 0->2
  vec2 clipSpace = zeroToTwo - 1.0; //convert from 0->2 to -1->+1 (clipspace)
  gl_Position = vec4(clipSpace, 0, 1); //covert the gl_Position to Clip space
  v_texCoord = vec2( a_texCoord.s, a_texCoord.t);//texCoord for Fragment Shader
} </script>
```

Listing 3. An example of Fragment Shader—changing equations due to alpha value for vertices, faces, normals and texture's coordinates.

```
<script id="fragmentShader2"type="x-shader/x-vertex">
uniform sampler2D u_image; //sampler2D allows to sample pixel colours
                              from a 2D texture
varying vec2 v_texCoord; // the texCoords from the Vertex Shader.

void main() {
  vec4 finalColor = texture2D(u_image, v_texCoord);
  / finalColor[i] - array of BGRA values, where i=0 is B channel,
  i = 1 is G, i = 2 is R, i = 3 is A /
  if (finalColor[3] == 0.50) { //Vertices, alpha = 128
   float float_n=float(finalColor[2]+(finalColor[1]*256.0+
              +finalColor[0])/10000.0); }
  else if (finalColor[3] == 0.40) { //Faces, alpha = 100
   float float_n = float(finalColor[2] + finalColor[1]*256.0); }
  else if (finalColor[3] == 0.20) { //Normals, alpha = 50
   float float_n = float((finalColor[0] + finalColor[1]*256.0)/10000.0); }
  else if (finalColor[3] == 0.10) { //UV's, alpha = 25
   float float_n = float((finalColor[0] + finalColor[1]*256.0)/10000.0); }
  gl_FragColor = encode_float(float_n * 255.0); //output for CPU }
</script>
```

Re 5. Outputs of rendering are saved in invisible canvas, read by *readPixels* WebGL function and forwarded to arrays, which are used to create vertex buffers for 3D scene.

Present implementation of WebGL standard does not support Geometry Shader, which could allow to change geometry of particular 3D data. The texture returned by the Fragment Shader contains of float values which are saved in invisible canvas, read by *readPixels* WebGL function and forwarded to arrays of floats, which are used to create vertex buffers for 3D scene.

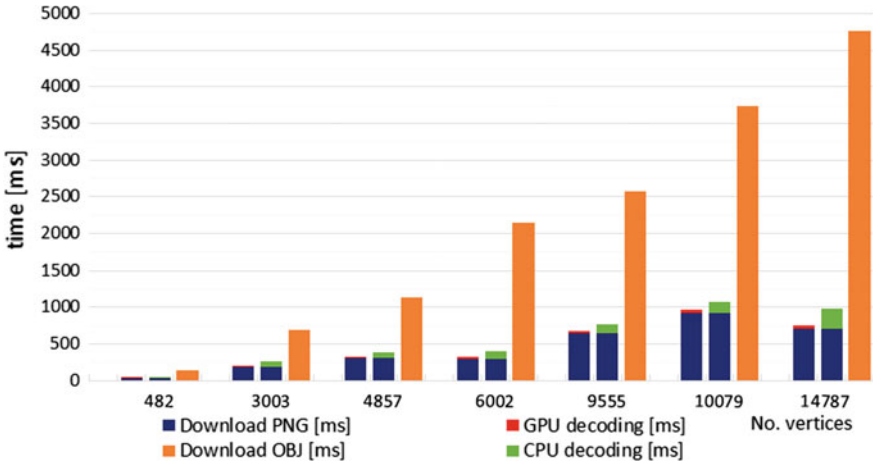


Fig. 6 Downloading (OBJ and PNG) and decoding (PNG) time comparison on CPU and GPU for different complexity of geometry

4 CPU and GPU Benchmarks

The authors' idea of saving geometry in PNG files is likely to decrease the size of files. On the other hand, there are additional calculations that need to be performed on client side during decoding process. Of course, it takes much less time to transfer smaller file via Web (Fig. 6.). For example, a model with 77 407 vertices saved in OBJ file (13 MB) downloads about 25 s, while model encoded by authors' algorithm in PNG (2.25 MB) only 4.5 s (Table 1). For the same PNG file, a period of time that is needed to decode a data set is also short (less than 1 s), but GPU decoding is nine times faster than CPU. As Table 1, Figs. 6, and 7 show, there is a big change between calculations performed on CPU and GPU.[1]

Concerning small files, CPU processing is a bit faster than GPU. It results in a constant time that is needed to create context, trigger shaders and allocate a memory on a GPU. In fact, there are bigger changes for more complex geometry. In case of CPU processing, time grows quadratically—for two times larger input files required time doubles. On the contrary, GPGPU's dependency between file's size and time is logarithmic like, what is caused by parallelism of tasks on huge data sets. There is no such growth of appropriate time like during traditional computing. What is more, a processing time is even the same but for ten times larger input picture.

[1]Technical specification: CPU Intel i7-4702HQ 2.2GHz, GPU Nvidia K110M 2GB.

Table 1 Downloading (4Mb/s) and decoding time comparison on CPU and GPU for different complexity of geometry saved in PNG and OBJ file

No. vertices	File's size (KB)		Downloading (ms)		Decoding (ms)	
	OBJ	PNG	OBJ	PNG	CPU	GPU
482	69	17	135	33	10	14
3003	355	93	693	182	75	23
4857	578	155	1129	303	89	22
6002	1102	150	2152	293	101	28
7872	1126	318	2199	621	117	37
9555	1321	328	2580	641	125	41
10079	1916	471	3742	920	146	43
14787	2442	363	4770	709	270	49
23029	3428	838	6695	1637	368	55
39202	6882	1377	13441	2689	688	74
77407	**12971**	**2308**	**25334**	**4508**	**936**	**102**
240002	21852	3595	42680	7021	1932	351
1120000	50812	7884	99242	15398	6891	533

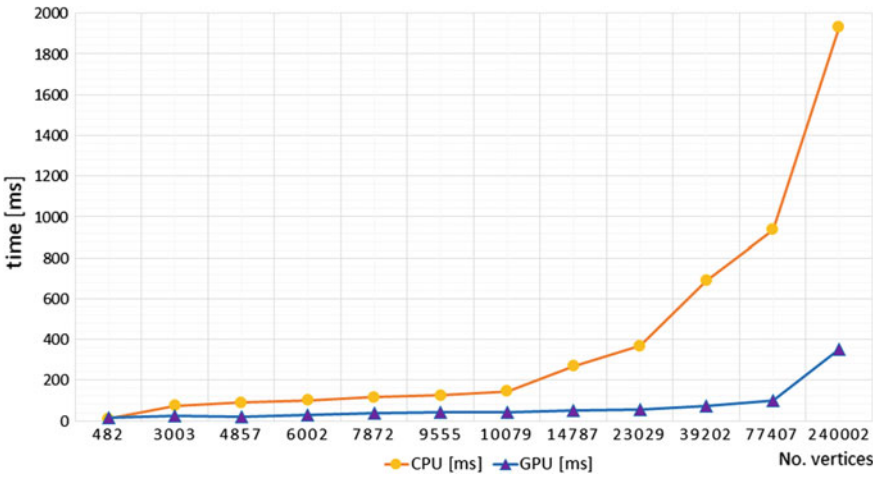


Fig. 7 Decoding time comparison on CPU and GPU for different complexity of geometry

5 Conclusions

The virtual 3D reconstructions using portable and crossplatform technologies are burdened with many restrictions. Now, most of modern 3D technologies are not standards and there are many aspects of improving them. Despite our proposed solutions are in development stage, they are promising so far. This fact can be confirmed by

achieved results—the reduction of file's size and time of decoding is over few times. There are many reasons to continue researches in this way, but Internet technologies (like WebGL) are demanding, needing many techniques of optimization to reach the main aim.

The analysis of reached results (Table 1, Figs. 6 and 7.) confirms that proposed solution highly reduces required time to decode data sets, what also makes it faster to create a geometry and display it on the screen. It is easy to notice that CPU processing has almost quadratic complexity, so obligatory time is strictly connected with the size of input data. On the other hand, there is no such dependency while GPU processing is performed—even when the input file is several times bigger, there is a change by only few percent. It is also worth to mention that processing time on CPU is shorter in case of really small sets (about few hundred of vertices), what is caused by constant time required to allocate memory and compile shaders while using GPU. Summing up, the application of GPGPU allowed to reduce appropriate time needed to decode a data encoded in PNG file even few times. What is more, it makes sense to carry on this kind of computing for relatively huge data sets, what is not a problem because of trends in modern tridimensional technologies.

Acknowledgments The research is partially performed within the project "Virtual reconstructions in transitional research environments—the Web portal: Palaces and Parks in former East Prussia" supported by Liebnitz Gemeinschaft in the years 2013–2016.

References

1. Behr, J., et al.: X3DOM: a DOM-based HTML5/X3D integration model. In: Proceedings of the 14th International Conference on Web3D Technology (Web3D 09), ACM, pp. 127–135 (2009)
2. Dworak, D., Pietruszka, M.: Fast Encoding of Huge 3D Data Sets in Lossless PNG Format. Advances in Intelligent Systems and Computing: New Research in Multimedia and Internet Systems, vol. 314, 15–24. Springer (2015)
3. Fung J., Mann S.: Computer Vision Signal Processing on Graphics Processing Units. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), Montreal, Quebec, Canada, pp. 93–96 (May 2004)
4. Goddeke D.: GPGPU Basic Math Tutorial (online: 15.04.2016). http://www.mathematik.uni-dortmund.de/~goeddeke/gpgpu/tutorial.html. Accessed 18 April 2016
5. https://github.com/KhronosGroup/glTF/wiki/converter. Accessed 18 April 2016
6. https://github.com/mrdoob/three.js/blob/master/utils/converters/obj/convert_obj_three.py. Acces
sed 18 April 2016
7. Khronos: Efficient, Interoperable Transmission of 3D Scenes and Models. https://www.khronos.org/gltf. Accessed 18 April 2016
8. Limper, M., Jung, Y., Sturm, T., Franke, T., Schwenk, K., Kuijper, A.: Fast, progressive loading of binary-encoded declarative -3D web content. In: Computer Graphics and Applications, IEEE, vol. 33, pp. 26–36 (2013)
9. Rost, J., Licea-Kane, B.: OpenGL Shading Language. 3rd edn. Addison-Wesley (2010)
10. Shreiner, D., Angel, E.: Interactive Computer Graphics with WebGL: Global Edition. Pearson Education (2014)
11. Sons K., et al.: XML3D: Interactive 3D graphics for the web. In: Proceedings of the 15th International Conference on Web3D Technology (Web3D 10), ACM, pp. 175–184 (2010)

12. Trevett, N.: 3D Transmission Format, NVIDIA (June 2013)

An Asymmetric Approach to Signature Matching

Tatiana Jaworska

Abstract The image signature concept can be a successful approach to image comparison in content-based retrieval, but it is a very challenging task. Fundamental for this purpose is defining signature similarity. There exist a lot of similarity models which measure similarity of images or their objects in multimedia databases. In order to deal with semantic retrieval, we have introduced a three stage search engine. At each stage, we need different but equally effective similarity measures. Here, we have analysed asymmetric and symmetric approaches to signature comparison. In our experiment, we present an extensive comparison of some similarity measures dedicated to image retrieval.

Keywords Similarity measure · Image signature · Search engine · CBIR

1 Introduction

In the recent years, many researchers have intensively analysed similarity evaluations between whole images, their fragments, or some image elements, such as contours. Content-based similarity models have been developed [1] so as to comply with the system needs and user requirements regarding semantic multimedia retrieval.

According to Beecks et al. [1], a similarity model between the query and image or a group of image objects can be determined, even for a large multimedia database, by working out only the distance between their corresponding feature representations. We claim that, even though a query is automatically generated by a CBIR system or is prepared manually by a user as, for instance, we proposed in our system introducing a special dedicated GUI [2], the introduction of the spatial object location is strongly recommended.

T. Jaworska (✉)
Systems Research Institute Polish Academy of Sciences,
01-447 6 Newelska Street, Warsaw, Poland
e-mail: Tatiana.Jaworska@ibspan.waw.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_3

Therefore, we provide in this paper a comparison between the Beecks' concept of feature signatures and their similarity measure, and our search engine concept where image signature and spatial object location are treated as global image information, but object features are only local. We decided to compare two kinds of signatures in order to check what gain it would bring if we found objects and compare their locations, as we proposed in our search engine [3, 4]. The aforementioned knowledge is crucial for the effectiveness of multimedia retrieval systems.

2 Signature Matching

2.1 Metrics Properties

Generally, when we analyse a metric space we assume by default that four basic conditions are satisfied:

- Non-negativity: $d(x, y) \geq 0$;
- Identity: $d(x, y) = 0 <=> x = y$;
- Symmetry: $d(x, y) = d(y, x)$; (1)
- Triangle inequality: $d(x, y) + d(y, z) \geq d(x, z)$ for any points x, y, z of the set.

These conditions express our common notions of distance. For example, the distance between distinct points is positive and the distance from point A to B is equal to the distance from B to A.

We may also need to find the distance between two vectors, namely, feature vectors. Then, in a normed vector space $(X, \|\bullet\|)$ we can define a metric on X by

$$d(x, y) = \|x - y\| \quad (2)$$

A metric defined in such a way is translation invariant and homogeneous. The most widely used similarity measure is the Euclidean measure. It can be applied to measure the distance between two points or between two feature vectors.

However, in real life the symmetry is questionable, for example, the way up a hill and down a hill takes different time. A similar situation is when we compare images. We can imagine various criteria, for instance, the number of particular elements or segments which constitute a query. Hence, when we select as a query an image among the previous matching images, we obtain a different set of matching images because the symmetry is incomplete.

In such a situation a quasimetric may be needed. A quasimetric is defined as a function that satisfies the previously mentioned axioms for a metric without symmetry:

$$d(x, y) \neq d(y, x). \tag{3}$$

This notion is more often used in practice than in mathematics, and that is why sometimes it is called a semimetrics [5].

2.2 Signatures

In our system [3], at the first stage, objects o_{ij} are extracted from an image I_i based on low-level features. These features are used for object classification. Additionally, the objects' mutual spatial relationship is calculated based on the centroid locations and angles between vectors connecting them, with an algorithm proposed by Chang and Wu [6] and later modified by Guru and Punitha [7], to determine the first three principal component vectors (PCV_{oi}, $i = 1, \ldots, 3$ for each object o_{ij}). Spatial object location in an image is used as the global feature [4].

Definition 2.1 (*Image signature* [3])
Let the query be an image I_q, such as $I_q = \{o_{q1}, o_{q2}, \ldots, o_{qn}\}$, where o_{ij} are objects. An image in the database is denoted as I_b, $I_b = \{o_{b1}, o_{b2}, \ldots, o_{bm}\}$. Let us assume that in the database there are, in total, M classes of the objects marked as L_1, $L_2, \ldots, L_M$. Then, as the image signature I_i we denote the following vector:

$$\mathrm{Signature}(I_i) = [\mathrm{nobc}_{i1}, \ \mathrm{nobc}_{i2}, \ \ldots, \mathrm{nobc}_{iM}] \tag{4}$$

where: nobc_{ik} are the number of objects o_{ij} of class L_k segmented from an image I_i. Note that the length of a signature is always the same and is equal to M.

As the second kind of signature we adopt the feature signature defined by Beecks et al. [8] in 2009 who aggregated features into a compact feature representation, for the purpose of effective calculation of similarity between two data objects. Rubner et al. [9], used two common feature representation types: feature histograms and feature signatures, which were worked out from global partitioning of the feature space and local feature clustering for each data object. Contrary to our approach, these authors applied global partitioning to the feature space, regardless of feature distributions of single objects, in order to create feature histograms which in turn correspond to the number of features located in the global partition.

According to Beecks et al. feature signature is defined as follows:

Definition 2.2 (*Feature signature* [1])
Let $FS \subseteq R^k$ be a feature space and $C = C_1, \ldots, C_k$ be a local clustering of the features $f_1, \ldots, f_n \in FS$ of object o_{ij}. Then a feature signature S^o of length M^i can be defined as a set of tuples from a $FS \times R^+$ such as:

$$S^o = \{c_k^o, w_k^o, k = 1, \ldots, M\}. \tag{5}$$

where: $c_k^o = \frac{\sum_{f \in c_k} f}{|c_k|}$ is a centroid of similar objects o_{ij} of image I_i and $w_k^o = \frac{|c_k|}{n}$ is its weight.

It means that a feature signature S^o of object o_{ij} is a set of centroids $c_k^o \in FS$ with the corresponding weights $w_k^o \in R^+$.

According to Definition 2.2. carrying out the feature clustering individually for each data object reflects aggregation of feature distribution in a better way than any feature histogram. However, feature histograms are a special case of feature signatures, whose centroids stay the same for the whole database and the information about objects is reflected only via weights, which results in a limitation of object representation.

By this approach, Beecks et al. aggregated the objects' location in the feature space which is substituted only by grouping similar feature values in signature and histogram form. They proposed only seven basic features: two coordinates, three components of colour and two texture descriptors, whereas we offered 45 features for a particular object, for example: moments of inertia and Zernike's moments [10].

In our adaptation of their method, a number of objects of a particular class were interpreted as weights. Object centroids represent locations of real, early segmented, objects in the image space. Here, class centroids are situated in the geometrical centre among particular object centroids. We also use different methods to determine the similarity in these two approaches.

2.3 Similarity Functions

Asymmetry is one of the most controversial properties of similarity. In this subsection we describe the asymmetric approach to image signature matching and a signature quadratic form distance in comparison with standard similarity measures, such as Euclidean, absolute difference or Hamming. All the measures are implemented in our search engine [3, 4].

In order to answer the query I_q, we compare it with each image I_b from the database in the following way. A query image is obtained from the GUI, where the user constructs their own image from selected DB objects. First of all, we determine a similarity measure $\mathrm{sim}_{\mathrm{sgn}}$ between the signatures of query I_q and image I_b:

$$\mathrm{sim}_{\mathrm{sgn}}(I_q, I_b) = \sum_i (\mathrm{nob}_{qi} - \mathrm{nob}_{bi}) \tag{6}$$

computing it as an equivalent with the Hamming distance between two vectors of their signatures (cf. (4)), such that $\mathrm{sim}_{\mathrm{sgn}} \geq 0$ and $\max_i(\mathrm{nob}_{qi} - \mathrm{nob}_{bi}) \leq \mathrm{thr}$, thr is the limitation of the quantity of elements of a particular class by which I_q and I_b can

differ. It means that we select from the DB images with the same classes as the query. The above comparison is asymmetric because if we interchange the query and the image, we obtain negative similarity value, that is: $\mathrm{sim}_{\mathrm{sgn}}(I_q, I_b) = -\mathrm{sim}_{\mathrm{sgn}}(I_b, I_q)$. Then, the condition of non-negativity of similarity is incomplete. This fact is crucial from the semantic matching point of view because the human brain recognizes things in context with others.

If the maximum component of (6) is bigger than a given threshold (a parameter of the search engine set by the user), then image I_b is discarded. Otherwise, it goes to the next step and we find the spatial similarity $\mathrm{sim}_{\mathrm{PCV}}$ (7) of images I_q and I_b, based on the City block distance between their PCVs as:

$$\mathrm{sim}_{\mathrm{PCV}}(I_q, I_b) = 1 - \sum_{i=1}^{3} |PCV_{bi} - PCV_{qi}| \tag{7}$$

Definition 2.3 (*Signature Quadratic Form Distance* [1])

If $S^o = \{c_k^o, w_k^o, k = 1, \ldots, M\}$ and $S^q = \{c_k^q, w_k^q, k = 1, \ldots, N\}$ are two feature signatures, then the Signature Quadratic Form Distance (SQFD) between S^o and S^q is defined as:

$$\mathrm{SQFD}(S^o, S^q) = \sqrt{(w_o| - w_q)A(w_o| - w_q)^T} \tag{8}$$

where: $A \in R^{(M+N)\times(M+N)}$ is the similarity matrix, $w_q = (w_1^q, \ldots, w_m^q)$ and $w_o = (w_1^o, \ldots, w_n^o)$ are weight vectors and $(w_o| - w_q) = (w_1^o, \ldots, w_n^o, -w_1^q, \ldots, -w_m^q)$.

The similarity matrix **A** can be constructed assuming that there exists a similarity function f: $FS \times FS \rightarrow R$. The $a_{k,l}$ components of **A** are calculated as follows:

$$a_{k,l} = f\left(c_k^o, c_l^q\right) = \frac{1}{1 + d\left(c_k^o, c_l^q\right)} = \frac{1}{1 + \left[\left(c_{k,x}^o - c_{l,x}^q\right)^2 + \left(c_{k,y}^o - c_{l,y}^q\right)^2\right]} \tag{9}$$

where: $k, l = 1, \ldots, N + M$.

In our approach, there is the same number of classes for each image signature ($N = M$), hence we decided to assume the length of vectors w_q and w_o equal to M which implies the size of a square matrix $\mathbf{A}_{[M\times M]}$. Then the signature form distance (8) can be simplified to the form:

$$\mathrm{SQFD}(S^o, S^q) = \sqrt{w_o A w_q^T} \tag{10}$$

and $a_{k,l}$ components are computed only for $k, l = 1, \ldots, M$, according to (9). Here, in Definition 2.3 and in our approach, S^q means a query signature, whereas S^o means image signatures in the database. The signature similarity, computed according to SQFD (cf. (10)), gives more information than the one computed it according to $\mathrm{sim}_{\mathrm{sgn}}$ (cf. (6)) which is seen in the results.

3 Results

Below we present examples of matching results obtained for the above-mentioned similarity measures. We applied two queries designed in our GUI. The former is a semidetached building with a hip roof and the latter is a terraced house with three gable roofs. From the semantic matching point of view, the best results should be images of houses with the same kinds of roofs and similar number of building segments. All figures present query (far left picture in each) and 11 best matched images which are ordered decreasingly, according to the similarity to the query. Figures 1 and 2 present results found according to the asymmetric image signature (cf. (6)). We can see that both results contain buildings with two flat roofs and one with a semicircular type, which are not desired.

Figures 3 and 4 present matching for both queries computed according to our modification of signature form distance (cf. (10)) and all results fulfil the semantic requirements. Figures 5 and 6 present matching for these same queries, computed according to the signature quadratic form distance (cf. (8)).

Generally, in the case of semantic comparison, it is difficult to compare these results in a quantitative way, so that is why we present the result in full form. This gives us the opportunity for a qualitative evaluation. Hence, as we can see, especially in Fig. 6, where there are no fully correct matchings because separate doors,

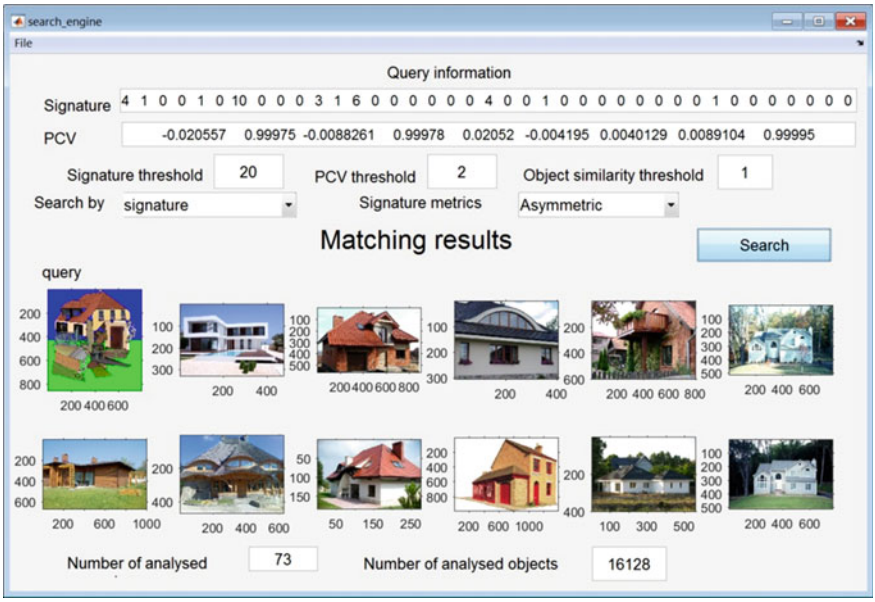


Fig. 1 Matching results for image signature for query 1

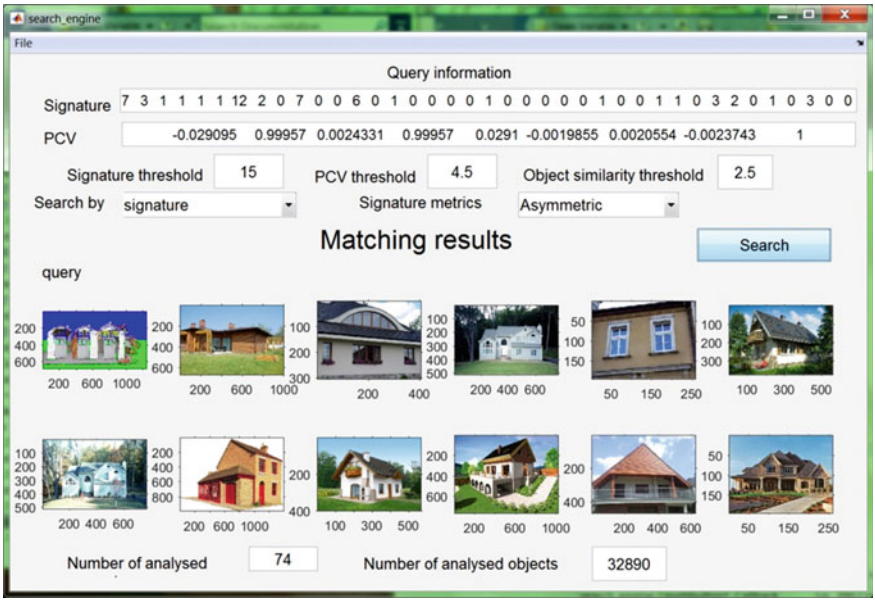


Fig. 2 Matching results for image signature for query 2

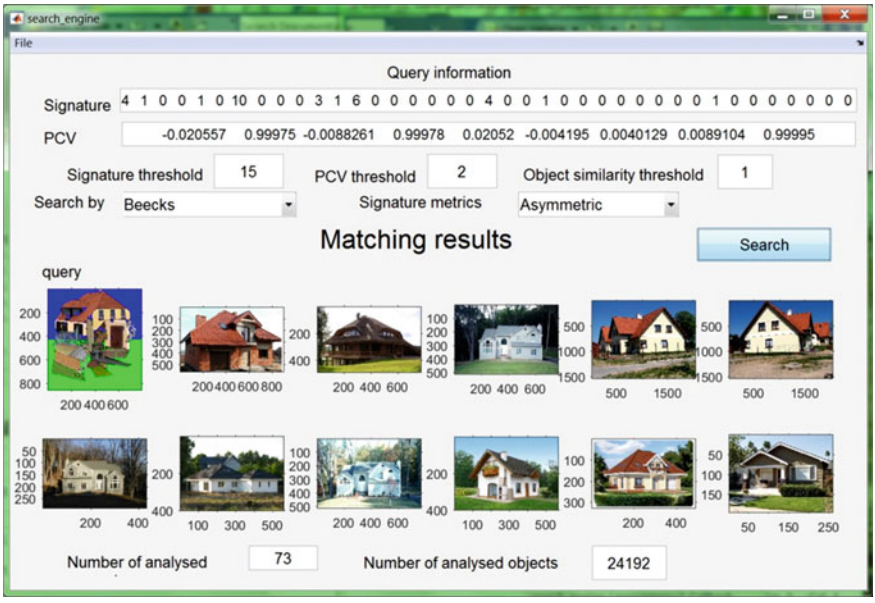


Fig. 3 Matching results for the signature form distance for query 1

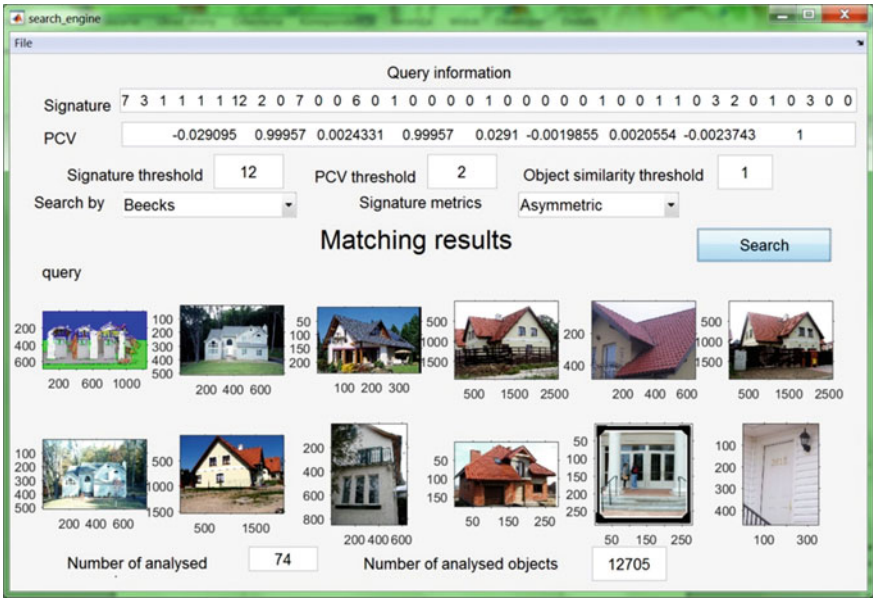


Fig. 4 Matching results for the signature form distance for query 2

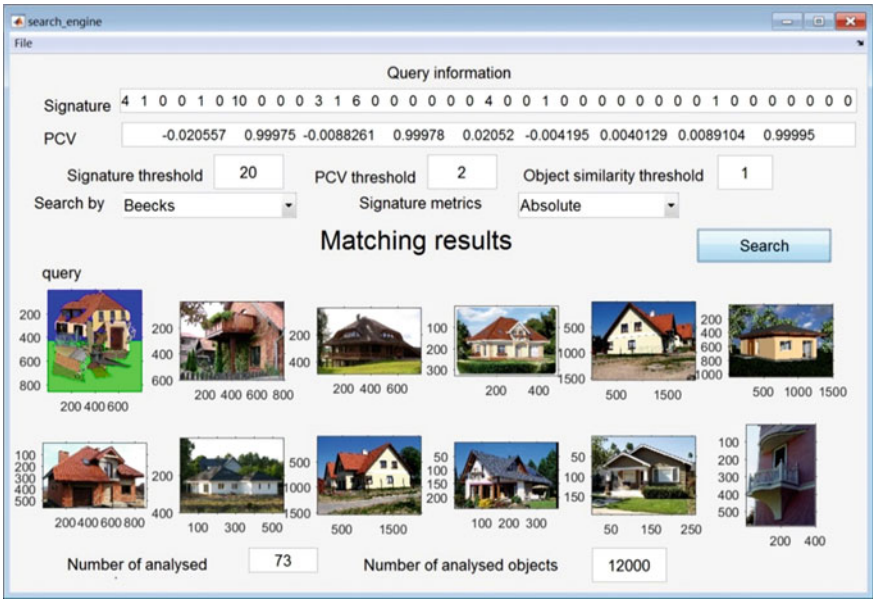


Fig. 5 Matching results for signature quadratic form distance for query 1

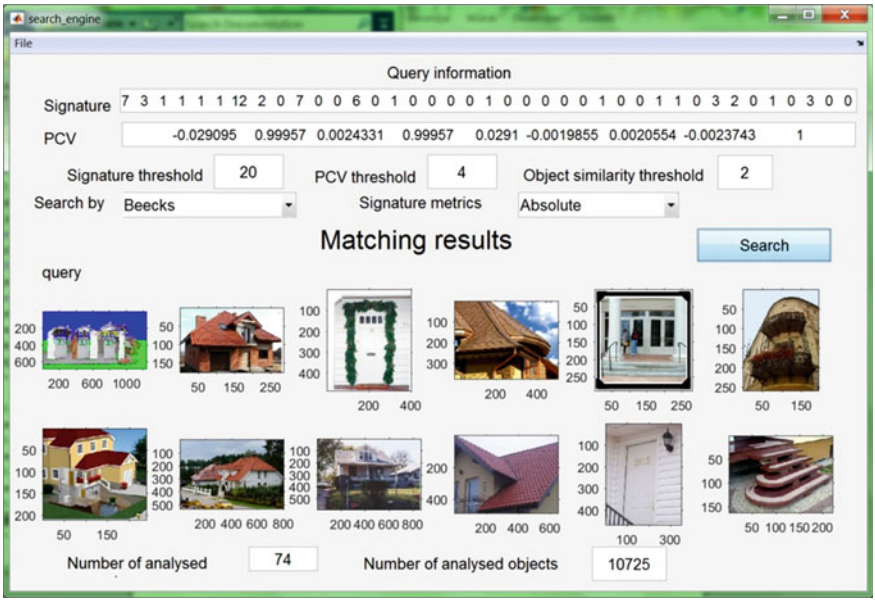


Fig. 6 Matching results for signature quadratic form distance for query 2

stairs, balconies appear strongly undesirable. It happens because the Beecks' team analysed less information about object spatial location than we did. Even though we used the simplified signature form distance, we obtained better results thanks to the fact that we added the separate object spatial location similarity (cf. (7)).

4 Discussion

In our analysis we have not decided to add such a popular approach as the SIFT method [11], because it mainly concentrates on finding a particular object similarity without a deep object spatial location analysis. The example of such a matching is shown in Fig. 7 where to the three terraced buildings seven houses with balconies were matched because a gutter or another less important element added in the query were found in the DB.

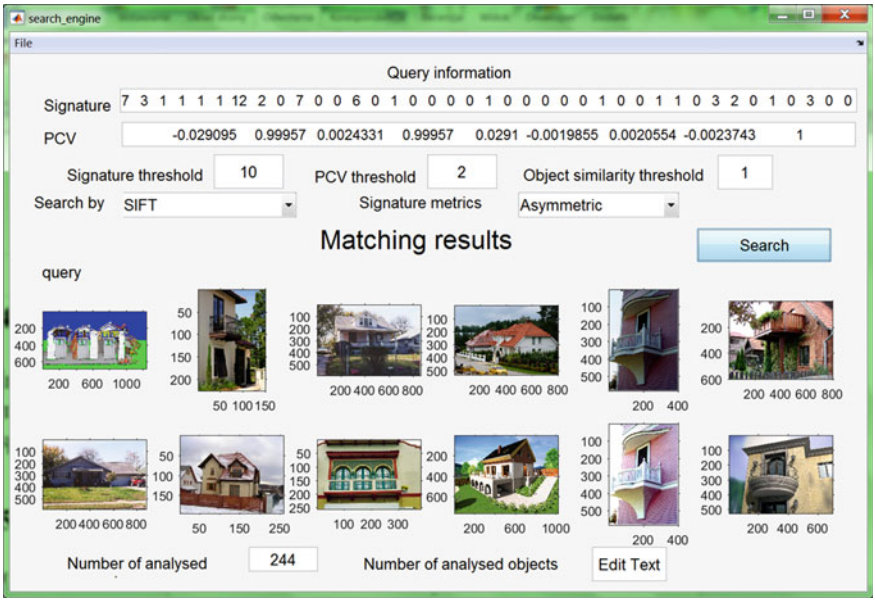


Fig. 7 Matching results based on the SIFT method for query 1

5 Conclusion

In this paper we compare the asymmetric and symmetric similarity measures applied to two kinds of signatures implemented in our content-based image retrieval system. In order to present the evaluation of the above-mentioned similarity measures, we used the database created in our institute containing mainly images of houses coming from the Internet.

We can observe that in a situation when a signature similarity is enhanced by object spatial location, the quality of semantic matching is better. All these similarity measures are applicable to signatures of different size and structure.

References

1. Beecks, C., Uysal, M.S. Seidl, T.: A comparative study of similarity measures for content-based multimedia retrieval. In: Multimedia and Expo (ICME), Suntec City, 19–23 July 2010
2. Jaworska, T.: Query techniques for CBIR. In: Andreasen, T., Christiansen, H., Kacprzyk, J., Larsen, H., Pasi, G., Pivert, O., De Tre, G., Vila, M.A., Yazici, A., Zadrożny, S. (eds.) Flexible Query Answering Systems, vol. 400, pp. 403–416. Springer, Cracow (2015)
3. Jaworska, T.: A search-engine concept based on multi-feature vectors and spatial relationship. In: Christiansen, H., De Tré, G., Yazici, A., Zadrożny, S., Larsen, H.L. (eds.) Flexible Query Answering Systems, vol. 7022, pp. 137–148. Springer, Ghent (2011)

4. Jaworska, T.: Spatial representation of object location for image matching in CBIR. In: Zgrzywa, A., Choroś, K., Siemiński, A. (eds.) New Research in Multimedia and Internet Systems, vol. 314, pp. 25–34. Springer, Wrocław (2014)
5. S. Rolewicz, Functional Analysis and Control Theory: Linear Systems, vol. Series: Mathematics and its applications, PWN-Polish Scientific Publishers, Warsaw 1987
6. Chang, C.-C., Wu, T.-C.: An exact match retrieval scheme based upon principal component analysis. Pattern Recogn. Lett. **16**, 465–470 (1995)
7. 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)
8. Beecks, C., Uysal, M.S. Seidl, T.: Signature quadratic form distances for content-based similarity. In: ACM Multimedia, Beijing, China, 19–24 Oct 2009
9. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vision **40**(2), 99–121 (2000)
10. Teague, M.R.: Image analysis via the general theory of moments. J. Opt. Soc. America **70**(8), 920–930 (1980)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60** (2), 91–110 (2004)

Automatic Playing Field Detection and Dominant Color Extraction in Sports Video Shots of Different View Types

Kazimierz Choroś

Abstract Sports videos are the most popular videos searched in the Web. To retrieve efficiently and effectively sports videos we need to use very sophisticated techniques of content-based analysis. Many strategies of content-based indexing have been already proposed, tested, and applied to categorize video shots in sports news videos. One of the techniques is based on player scene analyses leading to the detection of playing fields. The characteristic of a playing field strongly depends on the sports category. Some sports videos are characterized by a very dynamic background, others by a static background, close-up view of players, in-field medium view, wide view, or out of field view of the audience, small or great objects of foreground, homogeneous type of playing field with one dominant color or very diversified field. The recognition of such sports video features as dominant color of playing field and type of shot view can significantly help to categorize sports news videos. The paper discusses some aspects of the processes of automatic dominant color extraction and playing field detection basing on the experiences achieved during the experiments performed in the Automatic Video Indexer AVI.

Keywords Content-based video indexing · Sports news videos · View type · Video shots and scenes · Playing field detection · Dominant color extraction · Automatic video indexer AVI

1 Introduction

Videos are more and more frequently searched in the Web. Internet users prefer to watch short videos with current news than to read texts. The huge number of videos requires very efficient methods of content-based analysis for video indexing.

K. Choroś (✉)
Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: kazimierz.choros@pwr.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_4

Many strategies of content-based video indexing have been already proposed, tested, applied and many new methods are further developed [1] (comprehensive bibliography of 192 references), [2]. News videos and sports news videos are the most important part of video archives. Therefore, many of these video indexing methods are dedicated to news videos. News videos as well as sports news videos have a special structure. First of all these videos are edited videos [3], i.e. they have been created from parts chosen during the montage process from a collection of recorded videos. In the case of broadcast news these chosen video parts report different significant political, economic, social, or historical events as well as sports events recorded in different places and at different times. The analysis of such edited news videos is usually based on the video structure, therefore, the analysis of a news video includes the temporal segmentation process of a video and then the recognition of video category of a given structural video unit.

Video categories defined for news video shots usually reflect the specificity of the shot content, these are for example as defined in [4]: anchor shot, animation (intro), black frames, commercial, communication (static image of reporter), interview, map, report (stories), reporter, studio (discussion with a guest), synthetic (tables, charts, diagrams), and weather. The main problem is to detect anchorperson shots because the anchorman is the most frequently speaking person in news [5]. The anchorman is detected by the methods based on template matching, by identifying different specific properties of anchor shots, but also on temporal analyses of shots. It is observed that an anchorman shots are introduced several times during the news and they are distributed all along the video timeline. The interview shots i.e. shots with a reporter interviewing other people as well as political statement shots are also frequent in news and the are very similar to anchorman shots. It may lead to wrong identification of anchorman shots.

Whereas in the case of sports news videos such a categorization can be also applied. However, the detection of sports discipline reported in a shot is much more desired. Sports news videos usually present the most amazing sports highlights such as the best parts of games, matches, contests, tournaments, races, cups, etc. These highlights are for example goals, free kicks or penalties in soccer, basketball shot or slam dunk, tennis winner balls, bicycle race finishes, etc. Different sports disciplines have different the most exciting actions. The first criterion in sports video retrieval is a sports category because it is the most frequent and easiest search key for the user searching video information on sports events. But the video retrieval system can also optimize the retrieval procedure by applying weighted methods of sports news video indexing [6] taking into account not only which sports disciplines are reported in news but also how many sports events of a given category have been reported and how long these events have been presented in news. The news videos can be then retrieved in video retrieval systems on the basis of the category reported in news as well as of the weighted attractiveness of the videos.

There are many strategies of the analysis and categorization of news or sports news frames [7]. Their efficiency also depends on the type of camera views. If the sports shot categorization is based on player face detection close views are very useful whereas the long view shots are completely useless. On the other hand when

a playing field [8] or court lines [9] should be detected the analysis of close view shots presenting the player faces cannot help such detections, it can be achieved mainly when long view shots are examined. The conclusion is that the choice of categorization strategy should depend of a type of camera view.

The detection of playing field is a significant process for categorization of sports news shots. Generally the playing fields are detected in sports tracking systems to facilitate the tracking of soccer, volley, or tennis balls [10] as well as ice-hockey pucks [11]. But the detection and the recognition of the specificity of playing fields are also very useful for sports category identification of shots in sports news. Many of the methods of playing field detection are based on dominant color identification. It seems obvious that when detecting soccer shots playing field should be green whereas for winter sports categories the dominant color should be white. Unfortunately not all sports categories have so deterministic color of playing field as soccer has, for example the color of clay tennis courts is usually red because they are made of crushed brick, but hard courts are of any color. Similarly mainly because of advertising policy the colors of surfaces of sports halls can be strongly diversified. Nevertheless, the color of playfield is a significant indicator when sports shots are categorized.

The paper is organized as follows. The next section describes the main related work in the area of automatic detection of playing fields and field colors. The types of camera views of sports news shots and the methods of automatic detection of view types will be discussed in the third section. The fourth section presents the modified techniques of dominant color identification applied in the AVI Indexer. The final conclusions and the future research work areas are discussed in the last fifth section.

2 Related Work

Dominant color of the field detected in a sports video has been used for classification of shot of soccer and basketball. It was applied in the video indexing system [12] which was designed to detect soccer goal and to summarize soccer game in real-time as well as to segment basketball match into plays and breaks by skipping fouls, free throws, and time-out events. The dominant field color was defined by the average values of color components. For a certain number of frames color histograms were calculated and the peak for each histogram was found. Then, such a minimal interval for each histogram peak was set that the pixel counts were lower that 20 % of the histogram peak. For this interval the average color was calculated for each color component. Pixels of the frame were included to the field if the distance of its color to the average color was not greater than a given threshold value.

To improve the results of dominant color detection when long view shots are analyzed the frame can be segmented according to the so-called golden ratio (Fig. 1). The frame is divided according to the ratio 3 to 5 to 3 (which is close to the golden ratio) in both directions. It enables us to remove these parts of the frame which are very probable out of the field, mainly the upper segments of the frame.

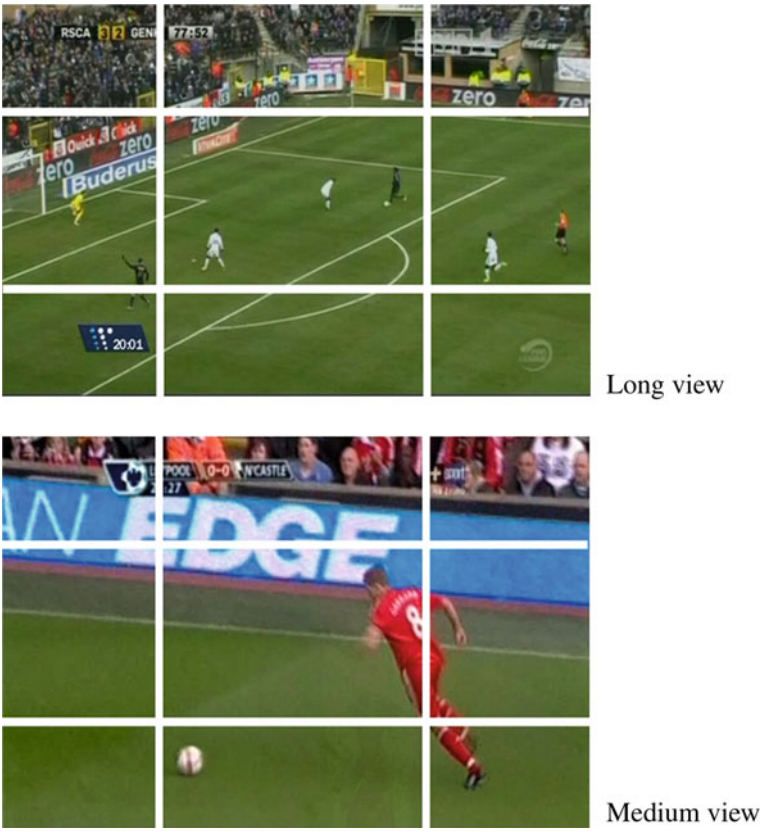


Fig. 1 Examples of soccer frames segmented according to golden ratio. The lower segments have greater values for playing field color detection and further analysis

In [13] the authors tested similar approach but for the playing fields with shadows. The dominant color of the soccer field is of course green, but because of differences resulting from natural or artificial stadium illumination the green color can differ and it can differ even in the same frame. Some parts of the playing field can be sunny, some other can be shadow areas. The authors analyzed the color histograms of hue and saturation in the HSV color space. A slight range centered the peak was adopted in the histogram identified as the dominant field color.

The method described in [14] proposed a solution that extracted the dominant color without any threshold setting by the use of so-called dominant sets clustering. The dominant sets clustering insures a principled measure of a cluster's cohesiveness as well as a measure of a vertex participation to each group. The first dominant set is the biggest one, next the other sets become smaller and smaller, therefore, the first dominant set corresponds to the dominant color in a frame. The first dominant set is supposed to be the dominant color.

In the player detection framework proposed for soccer videos and presented in [15] the top grass pixels are selected in the initially detected playing field, i.e. green points lying on highest column in the frame. Then, the linear least square method was used to estimate a quadratic curve from a point set. It leads to the detection of a more coherent area of a playing field.

A possible technique of the dominant color detection in the playfield is the use of the MPEG-7 dominant color descriptor (DCD). Such an approach has been used in a complex system of soccer video segmentation [16]. However, as it was observed, its results are not independent of illumination variability.

The ratio of dominant color in a frame, i.e. the ratio of the number of pixels belonging to the playing field and the total number of pixels in a frame is used to determine whether a frame belongs to a long view shot—if the ratio is high and we suppose that the most part of a frame presents playing field, a medium view shot—if the ratio has medium value, close-up view shot—if the ratio is relatively low, and finally a non-field view when the ratio is extremely low. Shot type depending on the camera view is an important information for sports shot categorization.

3 Diversity of Camera Views in Sports Videos

Camera views are very important parameter in video techniques and are very significant in surveillance systems or urban traffic control systems [17]. Sports shots can be also classified according to the camera settings. Different camera views are usually defined. The classification can depend on the content type. These are for example [18]: court/field views, players and coach, player close-up views, audience views, and setting long views. In sports news videos [19] there can be: intro, headlines, player shots, studio shots, and final animation.

For a given sports category shots can be classified on the basis of sports action reported in shots, such as: goals, corners, fouls, substitutions, and game play shots. A method to automatically classify shots into these classes for each frame of an edited football broadcast video has been proposed in [20]. Furthermore, an assumption has been formulated that the patterns of transitions between camera viewpoints are correlated with different important actions taking place during analyzed sports events.

Camera lens settings significantly differentiate shot types. The recorded shots can be: long views, middle views, close-up views, out of fields views [21]. Long views usually show us the global view of field presenting the great part of playing field and enabling the viewers to catch team tactics and to admire for example long passes in soccer games. A middle view presents several players in a zoom-in view and in most cases a whole body as well as a body movement of a player can be observed. Whereas close views or out of field views present only a part of body of players mainly faces or the small parts of audience to make possible the observation of the emotions of players, referees, coaches, or spectators.

As it was already mentioned the long views are much more adequate for such analyses as color playing field detection, calculation of number of players

participating in the game, detection of line specific for a given sports category but also for analysis of a game strategy.

The perceived distance between the camera and the players recorded in a shot can lead to the very detailed classification of sports shots as it has been defined in [8]. Fourteen different shot types have been proposed: close up shot head with simple background, close up shot head complex background, close up shot head mixture background, close up shot waist up simple background, close up shot waist up complex background, close up shot waist up mixture background, short distance shot presenting player(s) simple background, short distance shot presenting player(s) complex background, short distance shot presenting player(s) mixture background, short distance shot presenting spectators, long distance shot presenting centre of the field, long distance shot presenting right side of the field, long distance shot presenting left side of the field, and long distance shot presenting spectators.

The following observation is very important. Usually the dominant color does not change in a shot even when a camera quickly moves and rotates. The frequent rapid movements of the camera and objects (players) does not change the dominant color. During the whole soccer game the dominant color for all long views will be green except long views of audience.

The methods of the detection of player fields and players as well as their efficiency for the categorization for sports news videos has been tested in the AVI Indexer [22]. The detection of playing fields was useful to reduce the area where foreground objects, i.e. players in the case of sports videos, can be expected. In the first experiments two kinds of sports playing fields have been analyzed: soccer playing fields and tennis courts (Fig. 2). The most simple solution is to start by the

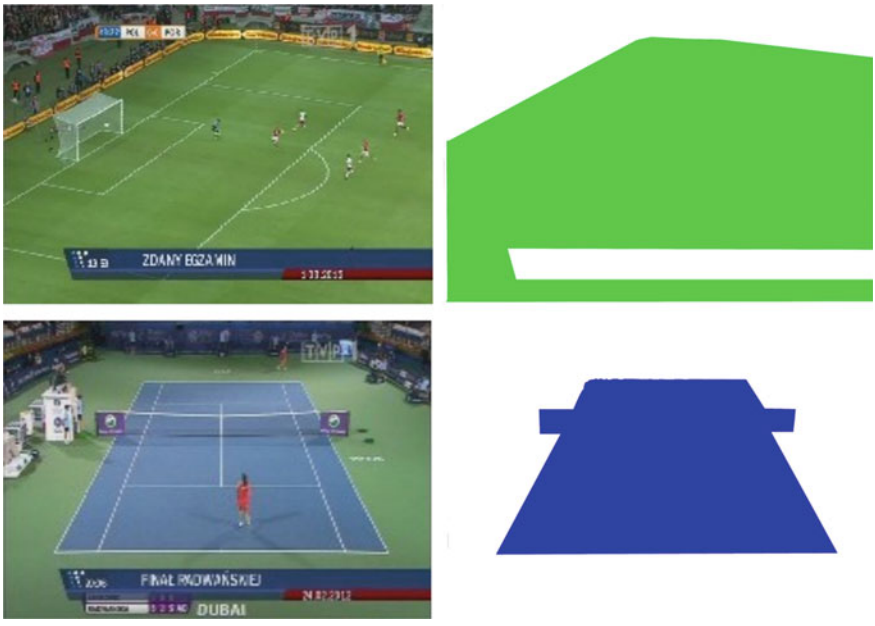


Fig. 2 Examples of results of soccer playing field detection and tennis court detection after removing all small areas of other than the dominant color

reduction of the number of colors in a analyzed frame to only eight basic colors. Then the dominant color is chosen. The dominant color is simply the most frequent color. To facilitate further analyses the region detected as the playing field is marked by the dominant color, moreover all small areas of other colors are removed. In the next step of the categorization process the players are detected in this area [23].

4 Playing Field and Player Detection in the AVI Indexer

The process of the detection of dominant color leading to the detection of playing fields is not always sufficiently efficient. The most favorable for the detection of playing field are camera long views of playing fields of colors significantly different from other parts of background. However, it happens during the sports event that camera is oriented on the end part of playing field for example in soccer games at the moment of corners. The camera records in such a case not only playing field but also an important part of audience which of course differs from the soccer playing field. Most frequently the techniques of dominant color detection have been proposed to one specific sports category. Many authors repeat obvious observation that the dominant color for long view shots of soccer games is green. Unfortunately, such techniques cannot be always applied for other sports categories like basketball or ice hockey.

The solution applied in the AVI Indexer is based on the method presented in [24]. In this paper, an algorithm of grass field detection has been described. As we know soccer field is green but however, a surface is not of uniform green color but is colored with a tone of green which varies depending on stadium, weather, and lightning conditions.

To avoid color variations resulting from different view angles, weather, it has been proposed to select the equally spaced frames and to compute the ratio of dominant color pixels in this set of frames. Then the two dimensional histograms in RG color space (two dimensional normalized version of RGB color space) are calculated for every frame and the peaks in the R and G histogram are determined. Such a procedure enables us to identify most of grass pixels but unfortunately there remain many noise pixels. The authors proposed a solution to improve the results of dominant color of playing field detection.

This technique is oriented not only on the detection of playing fields but also of objects in video frames. But the process of sports shot categorization based only on color of playing fields may not require to detect object and therefore the color space HSV can be applied. The main feature of a pixel in HSV color space is its hue (H). To determine the dominant hue in a shot some frames are chosen in a shot and then the HSV histograms are calculated. To significantly reduce the noise, for sports news frames these are mainly the parts of frames with audience, only lower 2/3 of the frame is analyzed. Audience is recorded in the vast majority of cases in the upper parts of the frames. For each selected frame we receive the proposed dominant color. But of course, it may happen that the dominant color for some frames is completely different than for others. To eliminate such cases significantly deviating

from the proper dominant color a threshold for acceptable standard deviation for H component is introduced. The proposed dominant H the most different from the average hue for all hues identified for selected frames is removed until the standard deviation is acceptable. The main procedure is used for S and V components eliminating the values significantly different than average values.

Finally, the HSV values of the dominant color are determined as the average values for filtered set of selected frames. A pixel will be included in the area of playing field in a sports news shot if the HSV components of this pixel are similar —three threshold should be set.

The unfavorable property of this procedure is that it can not be applied for winter sports shots where the dominant color is white (or similar to white). The white color can be also obtained independently of hue value when the saturation and brightness values are extremely height. If the dominant color is close to white color the detection should be done however in the RG color space.

Figure 3 presents a comparison of the results of dominant color detection in RG color space and HSV color space for the frames of different sports categories and for different camera views.

Fig. 3 Comparison of the results of dominant color detection in RG color space and HSV color space

5 Conclusions and Future Work

The general conclusion is that the sports shot categorization process requires an application of the most possible efficient techniques of automatic detections of camera view types of shots, automatic dominant color identifications, and finally automatic player fields detections. Although for different sports categories different categorization strategies are adequate, the detection of playing fields may help to avoid many ambiguities.

The characteristic of a playing field strongly depends on the sports category. Some sports videos are characterized by a very dynamic background, others by a static background, close-up view of players, in-field medium view, wide view, or out of field view of the audience, small or great objects of foreground, homogeneous type of playing field with one dominant color or very diversified field. The recognition of such sports video features as dominant color of playing field and type of shot view can significantly help to categorize sports news videos. The majority of already performed analyses were conducted with soccer videos. However, the categorization of sports news videos requires the efficient method of automatic detections of camera view types of shots, automatic dominant color identifications, and finally automatic player fields detections but for any sports category. The tests in the AVI Indexer have shown that the methods proposed for soccer videos with playing fields always green are not useful for winter sports categories with white dominant color or for sports categories without dominant color. This is the case for example for cycling races or marathon.

One of the promising directions of research is the analysis of sequences of shots and the recognition of structural video patterns specific for a given sports categories used in edited videos such as sports news videos.

References

1. Asghar, M.N., Hussain, F., Manton, R.: Video indexing: a survey. Int. J. Comput. Inform. Technol. **3**(1), 148–169 (2014)
2. Tien, M.C., Wu, J.L., Chu, W.T.: A comprehensive study of sports video analysis. Multimedia Analysis, Processing and Communications, Studies in Computational Intelligence 346, pp. 413–441. Springer, Berlin Heidelberg (2011)
3. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Event detection and recognition for semantic annotation of video. Multimedia Tools Appl. **51**(1), 279–302 (2011)
4. Valdés, V., Martínez, J.M.: On-line video abstract generation of multimedia news. Multimedia Tools Appl. **59**(3), 795–832 (2012)
5. Montagnuolo, M., Messina, A., Borgotallo, R.: Automatic segmentation, aggregation and indexing of multimodal news information from television and the Internet. Int. J. Inform. Stud. **1**(3), 200–211 (2010)
6. Choroś, K.: Weighted indexing of TV sports news videos. Multimedia Tools Appl. 1–20 (2015)
7. Choroś, K.: Video structure analysis for content-based indexing and categorisation of TV sports news. Int. J. Intell. Inf. Database Syst. **6**(5), 451–465 (2012)

8. Kapela, R., McGuinness, K., O'Connor, N.E.: Real-time field sports scene classification using colour and frequency space decompositions. J. Real-Time Image Process. 1–13 (2014)
9. Choroś, K.: Detection of tennis court lines for sport video categorization. Computational Collective Intelligence. Technologies and Applications. Springer, Berlin, Heidelberg, LNAI 7654, pp. 304–314 (2012)
10. Dang, B., Tran, A., Dinh, T., Dinh, T.: A real time player tracking system for broadcast tennis video. In: Intelligent Information and Database Systems ACIIDS'2010, LNCS 5991, pp. 105–113, Springer, Berlin Heidelberg (2010)
11. Yakut, M., Kehtarnavaz, N.: Ice-hockey puck detection and tracking for video highlighting. SIViP **10**(3), 527–533 (2016)
12. Ekin, A., Tekalp, A.M.: Shot type classification by dominant color for sports video segmentation and summarization. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'03, vol. 3, pp. 173–176 (2003)
13. Tong, X., Liu, Q., Lu, H.: Shot classification in broadcast soccer video. Electron. Lett. Comput. Vis. Image Anal. ELCVIA **7**(1), 16–25 (2008)
14. Li, L., Zhang, X., Hu, W., Li, W., Zhu, P.: Soccer video shot classification based on color characterization using dominant sets clustering. In: Proceedings of the Pacific-Rim Conference on Multimedia PCM, LNCS 5879, Springer, Berlin Heidelberg, pp. 923–929 (2009)
15. Tran, Q., Tran, A., Dinh, T.B., Duong, D.: Long-view player detection framework algorithm in broadcast soccer videos. In: Proceedings of the International Conference on Information and Computing ICIC, LNAI 6839, pp. 557–564, Springer, Berlin, Heidelberg (2011)
16. Maćkowiak, S.: Segmentation of football video broadcast. Int. J. Electron. Telecommun. **59** (1), 75–84 (2013)
17. Buch, N., Velastin, S.A., Orwell, J.: A review of computer vision techniques for the analysis of urban traffic. IEEE Trans. Intell. Transp. Syst. **12**(3), 920–939 (2011)
18. Duan, L.Y., Xu, M., Tian, Q., Xu, C.S., Jin, J.S.: A unified framework for semantic shot classification in sports video. IEEE Trans. Multimedia **7**(6), 1066–1083 (2005)
19. Choroś, K.: Automatic detection of headlines in temporally aggregated TV sports news videos. In: Proceedings of the 8th International Symposium on Image and Signal Processing and Analysis ISPA'2013, pp. 140–145 (2013)
20. Sharma, R.A., Gandhi, V., Chari, V., Jawahar, C.V.: Automatic analysis of broadcast football videos using contextual priors. Signal Image Video Process. 1–8 (2016)
21. Lang, C., Xu, D., Jiang, Y.: Shot type classification in sports video based on visual attention. In: Proceedings of the International Conference on Computational Intelligence and Natural Computing CINC'2009, IEEE, vol. 1, pp. 336–339 (2009)
22. Choroś, K.: Video structure analysis and content-based indexing in the Automatic Video Indexer AVI. In: Advances in Multimedia and Network Information System Technologies, Advances in Intelligent and Soft Computing, AISC 80, Springer, Berlin, Heidelberg, pp. 79–90 (2010)
23. Choroś, K.: Improved video scene detection using player detection methods in temporally aggregated TV sports news. In: Computational Collective Intelligence. Technologies and Applications, LNAI 8733, pp. 633–643, Springer, Switzerland (2014)
24. Nguyen, N., Yoshitaka, A.: Shot type and replay detection for soccer video parsing. In: Proceedings of the International Symposium on Multimedia ISM, IEEE, pp. 344–347 (2012)

Multi-label Classification with Label Correlations of Multimedia Datasets

Kinga Glinka and Danuta Zakrzewska

Abstract In multi-label classification tasks, very often labels are correlated and to not lose important information, methods should take into account existing dependencies. Such situation especially takes place in the case of multimedia datasets. In the paper, universal problem transformation methods providing for label correlations are considered. The comparison is done for proposed by authors Labels Chain technique [4] and well known methods which also take into account label correlations, such as Label Power-set, Classifier Chains and Ensembles of Classifier Chains. The performance of the methods is examined by experiments done on image, musical, audio and text datasets.

Keywords Multi-label classification · Labels chain · Label correlations · Multimedia datasets

1 Introduction

Multi-label classification is the task of predicting more than one predefined class. However, labels are correlated with each other and methods, do not taking into account dependencies between labels, lose important information, which could help in ameliorating classification effectiveness. Such situation especially takes place in the case of multimedia datasets. In the paper, universal problem transformation methods providing for label correlations will be considered. There will be compared the performance of the well known techniques and Labels Chain (LC) method. LC was firstly introduced in [4], and the method showed good performance on multidimensional datasets with the number of attributes significantly bigger than the number of

K. Glinka (✉) · D. Zakrzewska
Institute of Information Technology,
Lodz University of Technology, Wólczańska 215, Lodz, Poland
e-mail: 800559@edu.p.lodz.pl

D. Zakrzewska
e-mail: danuta.zakrzewska@p.lodz.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_5

instances. The experiment results were compared with the ones obtained by using Binary Relevance (BR) technique [18]. BR converts multi-label problem into several binary classification tasks by using one-against-all strategy. In spite of LC, BR ignores dependencies between labels. The experiments, presented in [4], showed that LC outperformed BR method on the considered datasets.

In the current research effectiveness of LC technique for multimedia datasets is checked comparing to the methods which also take into account label correlations. We consider: Label Power-set (LP) [1], Classifier Chains (CC) [13] and Ensembles of Classifier Chains (ECC) [14]. The performance of the methods is examined by experiments done on image, musical, audio and text datasets. The techniques are compared taking into account such measures as Exact Match, Accuracy, Precision and Recall as well as Hamming Loss and 0/1 Loss.

The remainder of the paper is organized as follows. In the next section, related work concerning multi-label classification methods exploiting label dependencies is presented. Then, the considered methods, evaluation indicators as well as datasets are described. In the following section, experiment results are discussed. Finally, concluding remarks and future research are depicted.

2 Related Work

Many authors considered label correlations in multi-label classification process. Researchers showed that making use of label dependencies can ameliorate the effectiveness of the classifiers (see [3] for example). Different approaches to multi-label classification with label correlations have been proposed so far. Most of them have been dedicated to multimedia datasets. Huang et al. [7] proposed the framework called Group Sensitive Classifier Chains (GSCC). They assumed that similar objects share the same label correlations and tend to have similar labels. GSCC firstly expands the feature space of label space and then, cluster them. Classifier chains are built on each cluster taking into account group specific label dependency graph. Active learning approach was investigated for image classification in [21, 22]. Ye et al. used cosine similarity to evaluate the correlations between the labels. Then, they develop an active learning selection strategy based on correlations to select label pairs for labeling [21]. Zhang et al. [22], in turn, proposed a high-order correlation driven active learning approach. They indicated that additionally to pair-wise label correlations, high-order correlation is also informative. Association rule mining was adopted to discover informative label correlation [22]. Both of the active learning approaches has been evaluated by experiments done on several datasets.

There exist many multi-label classification with label correlation methods, which are not dedicated to multimedia datasets but can be used for classification of that kind of data. In the next section there will be presented: Label Power-set (LP), Classifier Chains (CC) and Ensembles of Classifier Chains (ECC).

3 Materials and Methods

3.1 Methodology

There are considered four problem transformation methods that take into account label correlations: LP, CC, ECC and LC. All of them transform multi-label problems into one or more traditional single-label tasks [18], but additionally, they also map dependencies between labels. Let us consider the set of labels L and let K denotes a set of labels relevant to the instance.

Label Power-set (LP) creates new classes from all unique sets of labels existing in the training dataset. This technique allows to transform every complex multi-label problem into one single-label classification. Therefore, this method can be used regardless of number and variety of labels assigned to the instance. The main disadvantage of this approach is that it may lead to dataset with a large number of classes and few instances representing them. What is more, the method has tendency to overfit the training dataset because it can model only sets of labels observed in the training data. More details together with experiment results can be found in [1, 10, 18].

Classifier Chains (CC) is an improvement of Binary Relevance method (see [10, 18]) that model label correlations. Similarly to Binary Relevance approach, CC involves $|L|$ binary classifiers, but it includes all previous predictions as feature attributes. Classifiers are linked along a chain where each classifier deals with label l. The feature attributes space of each link in the chain is extended with the 0/1 label associations of all previous links. Thus, every following binary classifier for label l has one attribute more than the previous one (for l-1 label), that indicates if the instance has l-1 label or not. As the final classification result for a new instance, the method outputs the union of labels from all $|L|$ binary classifier results. Such chaining algorithm passes label information between classifiers, allowing *Classifier Chains* to take into account label correlations and thus, overcoming the label independence problem of Binary Relevance. The *Classifier Chains* method was investigated in [13, 14].

Ensembles of Classifier Chains (ECC) solves main problem of "chain order" of *Classifier Chains* method. It uses different random chain ordering for each iteration. Hence, each classifier model is likely to be unique and able to give different multi-label predictions. These predictions are summed by label so that each label receives a number of votes. A threshold is used to select the most popular labels which create the final predicted multi-label set. Ensembles are expected to provide better accuracy and overcome over-fitting. More details of the method can be found in [14].

Labels Chain (LC) was firstly introduced by authors in [4]. This method requires to learn $|K|$ multi-class classifiers, and consecutively uses result labels as new attributes in the following classification process. Thus, the labels chain is created, similarly to *Classifier Chains* technique, however, this method involves less classifiers. As classifications are not totally independent from themselves, such approach enables providing better predictive accuracy. The fact should be observed especially in multi-label problems with small number of labels in L, because in these cases

the value of a new, added attribute is more significant for classification process. The algorithm assumes the number of labels for instances to be known and can be also applied taking into account different order of classifications, with $|K|!$ available order combinations. Usually all possible label orders are considered and the best values are taken into account from all the single classifications. The detailed description and some experiment results of *Labels Chain* are presented in [4].

The performance of the considered methods is evaluated by using several multi-label example-based evaluation measures, described in [18]. They are based on the average differences of the true and the predicted sets of labels over all examples of the evaluation dataset. Let Y_i and Z_i be the compared sets—respectively, the set of true and predicted labels during classification process.

Classification Accuracy (also known as *Subset Accuracy* or *Exact Match*) [8, 23] is the most strict evaluation metric for multi-label classification. Contrarily to other measures, it ignores partially correct sets of labels, marking them as incorrect predictions, and requires all labels to be an exact match of the true set of labels. *Accuracy*, *Precision* and *Recall* are less restrictive metrics used for evaluating classification results, they take into account partially correct predicted labels. All the measures are defined as follows [18]:

$$Class\ Accur = \frac{1}{N}\sum_{i=1}^{N} I\left(Z_i = Y_i\right) \tag{1}$$

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|Y_i \cap Z_i\right|}{\left|Y_i \cup Z_i\right|} \tag{2}$$

$$Precision = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|Y_i \cap Z_i\right|}{\left|Z_i\right|} \tag{3}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|Y_i \cap Z_i\right|}{\left|Y_i\right|} \tag{4}$$

where: $I(true) = 1$, $I(false) = 0$, N is total number of instances in the test set.

There are also considered *Hamming Loss* and *0/1 Loss*. As they represent loss functions, their smaller values are connected with the better effectiveness of the algorithm. *Hamming Loss* [15] calculates the fraction of incorrectly classified single labels to the total number of labels, while *0/1 Loss* is the opposite of *Classification Accuracy*. The measures are defined as follows:

$$Hamming\ Loss = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|Y_i \triangle Z_i\right|}{|L|} \tag{5}$$

$$0/1\ Loss = 1 - \frac{1}{N}\sum_{i=1}^{N} I\left(Z_i = Y_i\right) \tag{6}$$

where $\triangle$ stands for the symmetric difference between two sets (set-theoretic equivalent of the XOR operation in Boolean logic) and L is the set of all labels.

3.2 Datasets

The considered label correlations-based methods are evaluated and compared on six multimedia datasets: image datasets *flags* and *scene*, audio dataset *birds*, musical dataset *emotions*, and text datasets *bibtex* and *medical*. All collections come from repositories of an open-source Java Mulan library [11].

The *flags* dataset contains 194 images of flags, described by 19 attributes and associated with one or more from 7 labels connected with colors on flag: red, green, blue, yellow, white, black and orange [5]. In next, *scene* dataset, each image is annotated with up to 6 labels as beach, sunset, fall foliage, field, mountain and urban, and described by 294 visual numeric features (spatial color moments in Luv color space) [1]. The audio *birds* dataset is the collection of 645 10-s audio recordings of birds, collected in the H. J. Andrews (HJA) Long-Term Experimental Research Forest. Each audio, characterized by 260 attributes, is paired with a set of species (from 19 classes) that are presented in [2]. Next, musical *emotions* dataset contains 593 songs, described by 6 clusters of music emotions (amazed-surprised, happy-pleased, relaxing-calm, quiet-still, sad-lonely and angry-aggressive), which are constructed based on the Tellegen-Watson-Clark model [17]. The *bibtex* dataset is a collection containing a large number of BibTex files downloaded from the internet. Each text is expressed with 1836 attributes and related up to 159 tags [9]. The last text dataset *medical* is created from 978 clinical free text reports from Computational Medicine Centers 2007 Medical Natural Language Processing Challenge, and each diagnostic report is associated to one or more disease code from 45 classes [12].

From all the datasets subsets of instances with 2, 3 and 4 relevant labels (depending on the set) have been selected, and respectively named as *flags2*, *flags3* and *flags4* for source dataset *flags*. All the characteristics of these datasets, which are used during experiments, such as the number of instances and attributes or label properties, are presented in Table 1.

It is worth to mention that number of occurring label orders in every dataset is much smaller than number of possible orders, calculated as combinations without repetition (Table 1). What is more, some combinations of relevant labels occur more often than others, and some do not appear at all. For example, in *flags2* dataset *blue* label is combined only with *yellow* and *white* classes, while *orange* occurs only with *white* label (Table 2). Top combination is *red* and *white* with 15 out of 43 instances. These observations allow to state that some label correlations exist in real datasets and that they should be used during classification process.

Table 1 Datasets characteristics, grouped by domain

Dataset	Instan.	Attrib.	Label properties		
			Num.	Orders[a]	Top combination of labels[b]
Images					
Flags2	43	19	7	9/21	Red, white (15)
Flags3	76	19	7	18/35	Red, blue, white (27)
Flags4	44	19	7	16/35	Red, green, white, black (8)
Scene2	176	294	6	8/15	Field, mountain (75)
Music/audio					
Birds2	101	260	19	49/191	Pacific wren, Swainson's thrush (9)
Birds3	67	260	19	46/969	Varied thrush, Swainson's thrush, golden crowned kinglet (7)
Emotions2	315	72	6	13/15	Amazed-surprised, angry-aggressive (81)
Emotions3	100	72	6	8/20	Relaxing-calm, quiet-still, sad-lonely (67)
Text					
Bibtex2	1840	1836	159	694/12561	TAG_evolution, TAG_software (72)
Bibtex3	1306	1836	159	777/657359	TAG_apob, TAG_epitope, TAG_mapping (58)
Medical2	212	1449	45	57/990	Class-4-753_0, Class-32-486 (76)
Medical3	14	1449	45	7/14190	Class-4-753_0, Class-32-486, Class-44-786_07 (4)

[a]Occurring label orders out of all possible orders
[b]Most often occurring combination of relevant labels (number of instances with this combination)

4 Experiment Results and Discussion

An experimental evaluation based on 12 multimedia datasets from image, audio/music and text domains, described in Sect. 3, was performed. The experiments were carried out to justify the value of *Labels Chain* method by comparing it to related techniques, which also take into account label correlations: *Label Power-set*, *Classifier Chains* and *Ensembles of Classifier Chains*. The algorithms were tested using the Naive Bayes classifier [6, 20] as the base classifier (compare [16]), and 10-fold cross-validation. During experiments Classification Accuracy, Accuracy, Precision, Recall, Hamming Loss and 0/1 Loss measures were calculated and compared. The software implemented for tests was based on Weka Software [19] and open-source Java Mulan library [11].

Table 2 Dataset *flags2*—distribution of combinations of relevant labels

	Red	Green	Blue	Yellow	White	Black	Orange
Red	X	3	0	5	15	1	0
Green	3	X	0	1	4	0	0
Blue	0	0	X	1	12	0	0
Yellow	5	1	1	X	0	0	0
White	15	4	12	0	X	0	1
Black	1	0	0	0	0	X	0
Orange	0	0	0	0	1	0	X

Table 3 *Classification accuracy* (%) and *Accuracy* (%) of Label Power-set (LP), Classifier Chains (CC), Ensemble of Classifier Chains (ECC) and Labels Chain (LC)

Dataset	Classification accuracy				Accuracy			
	LP	CC	ECC	LC	LP	CC	ECC	LC
Flags2	28.00	27.50	21.50	**30.23**	48.17	**49.71**	47.00	38.84
Flags3	**20.18**	20.00	20.00	13.16	**48.45**	42.87	45.10	36.99
Flags4	21.00	9.50	14.00	**22.73**	**61.07**	58.22	58.49	57.66
Scene2	57.35	40.85	40.85	**58.52**	**68.38**	66.73	66.43	59.18
Birds2	11.00	1.00	2.00	**11.88**	**24.55**	20.86	23.11	21.88
Birds3	**17.38**	10.24	13.10	7.46	38.07	38.94	**38.96**	31.54
Emotions2	**36.45**	24.74	23.79	33.33	53.18	**57.06**	57.02	43.33
Emotions3	68.00	68.00	63.00	**70.00**	**78.50**	78.10	78.47	78.15
Bibtex2	14.46	7.23	5.16	**21.96**	22.36	21.09	18.08	**26.34**
Bibtex3	7.03	0.84	0.99	**11.54**	13.94	21.46	18.77	**26.95**
Medical2	35.84	35.84	35.37	**55.19**	41.04	41.49	41.43	**58.79**
Medical3	3.00	0.00	0.00	**42.86**	56.50	53.75	51.83	**68.75**

Tables 3, 4 and 5 present evaluation measure values for all the considered datasets: *flags2*, *flags3*, *flags4*, *scene2*, *birds2*, *birds3*, *emotions2*, *emotions3*, *bibtex2*, *bibtex3*, *medical2* and *medical3*. For each collection bold values are the best ones, obtained for the particular evaluation measure.

Performance of the considered methods in terms of Classification Accuracy and Accuracy is shown in Table 3. For 9 out of 12 cases the best Classification Accuracy values were obtained for Labels Chain (LC) method with the difference between other methods up to even 40 %. Multi-label problems are very often expected to predict all relevant labels properly. Thus, this metric is considered as the most important one because it calculates only fully correct predictions. The results were summarized and compared in Fig. 1. Taking into account Accuracy, Label Power-set (LP) and Labels Chain (LC) techniques proved to be the most effective—respectively, with values of 68.75 % for *medical3* and LC and 68.38 % for *scene2* and LP method.

Table 4 *Precision* (%) and *Recall* (%) of Label Power-set (LP), Classifier Chains (CC), Ensemble of Classifier Chains (ECC) and Labels Chain (LC)

Dataset	Precision				Recall			
	LP	CC	ECC	LC	LP	CC	ECC	LC
Flags2	58.25	**59.83**	57.00	57.32	58.25	**61.50**	61.00	54.65
Flags3	**60.30**	51.01	54.58	56.46	**60.30**	55.06	57.68	51.75
Flags4	73.38	**76.93**	73.63	73.56	**73.38**	67.25	69.38	72.73
Scene2	73.89	**74.97**	74.07	74.57	73.89	77.55	**77.57**	74.15
Birds2	31.32	30.11	34.79	**36.18**	31.32	**39.23**	36.27	35.64
Birds3	47.86	55.44	**57.30**	49.21	47.86	**51.03**	47.62	46.77
Emotions2	61.54	62.08	62.24	**62.84**	61.54	75.33	**75.83**	58.25
Eemotions3	83.00	82.08	82.50	**84.01**	83.00	82.33	**85.33**	82.33
Bibtex2	26.30	27.87	26.85	**41.89**	26.30	34.54	26.22	**41.52**
Bibtex3	17.72	34.84	36.68	**44.67**	17.72	39.24	29.08	**40.51**
Medical2	43.64	45.17	45.76	**74.76**	43.64	44.11	43.16	**73.35**
Medical3	66.67	82.50	78.33	**84.62**	66.67	55.00	55.00	**78.57**

Table 5 *Hamming Loss* (%) and *0/1 Loss* (%) of Label Power-set (LP), Classifier Chains (CC), Ensemble of Classifier Chains (ECC) and Labels Chain (LC)

Dataset	Hamming loss				0/1 Loss			
	LP	CC	ECC	LC	LP	CC	ECC	LC
Flags3	**23.86**	25.07	25.36	24.58	72.00	72.50	78.50	**69.77**
Flags3	**34.03**	43.70	39.87	37.78	**79.82**	80.00	80.00	86.84
Flags4	**30.43**	30.79	32.00	30.52	79.00	90.50	86.00	**77.27**
Scene2	17.41	17.06	17.42	**17.05**	42.65	59.15	59.15	**41.48**
Birds2	14.46	18.72	14.98	**13.39**	89.00	99.00	98.00	**88.12**
Birds3	16.47	15.61	**14.96**	31.54	**82.62**	89.76	86.90	92.54
Emotions2	25.64	24.74	**24.68**	25.40	**63.55**	75.26	76.21	66.67
Emotions3	17.00	17.50	17.17	**16.67**	32.00	32.00	37.00	**30.00**
Bibtex2	1.85	3.85	3.04	**1.46**	85.54	92.77	94.84	**78.04**
Bibtex3	3.10	4.20	3.21	**2.07**	92.97	99.16	99.10	**88.46**
Medical2	5.01	2.88	2.83	**2.29**	64.16	64.16	64.63	**44.81**
Medical3	4.44	3.78	4.22	**2.38**	97.00	100.00	100.00	**57.14**

Table 4 presents Precision and Recall results. Evaluating by Precision the LC method performed the best for 3 out of 4 datasets from music/audio domain and for all text collections. It means that predictions of this technique are more exact than the others for these datasets.

The highest Precision is connected with the smallest number of non-relevant labels incorrectly predicted as relevant (small number of false positive results). In

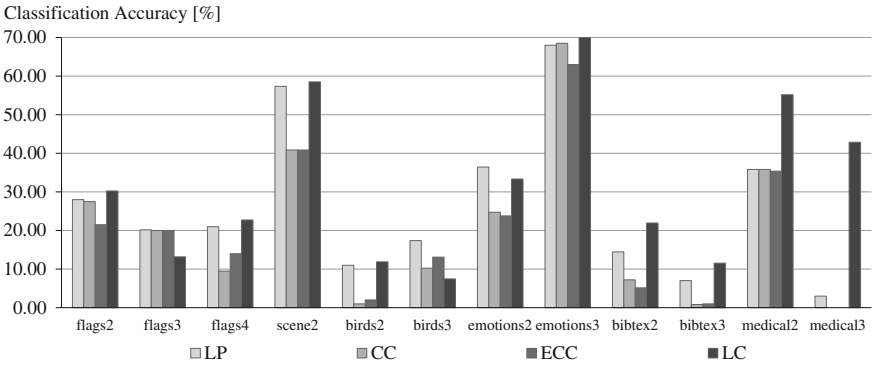


Fig. 1 Performance of Label Power-set (LP), Classifier Chains (CC), Ensemble of Classifier Chains (ECC) and Labels Chain (LC) in terms of *Classification Accuracy* (%)

terms of Recall, the results were more differentiated and no method outperforms the others. However, all the best results for text datasets (the smallest values of false negatives) were obtained by LC technique. For 3 music/audio datasets with the highest Precision for LC, obtained Recall is lower—some relevant labels could be left out during classification process in these cases.

Hamming Loss and 0/1 Loss measures are placed in Table 5. Their smaller values mean the better performance of the multi-label classifier. The best performance in terms of Hamming Loss, calculating incorrectly classified single labels, is observed for 7 out of 12 datasets classified by LC, especially the text ones. However, for collections with flag images, LP technique was indicated as the best one. Also results of 0/1 Loss confirm effectiveness of LC technique for almost all the datasets, except *flags3*, *birds3* and *emotions2*. This measure indicates predictions that are not fully correct, thus, it is also important metric for restrictive multi-label classification.

Summing up, most of the evaluation measures indicated LC technique as the best one for the considered datasets. It is easy to notice that the results for the most strict measure, Classification Accuracy, in most of the cases voted for Labels Chain method. This measure is especially important for multi-label problems which very often require all predicted labels to be an exact match of the true labels. What is more, this technique allowed to obtain the highest effectiveness of all the metrics for the text datasets. The advantage over the other methods was even equal to 40 % for *medical3* collection, while other techniques did not coped with it at all. On the other hand, rest of the methods had more partially correct predictions. The considered loss functions also showed the best performance of LC method for most of the datasets.

5 Conclusions

In the paper, multi-label classification methods which provide for label correlations are evaluated and compared, taking into account several evaluation measures including two loss functions. Effectiveness of Labels Chain technique is checked for multimedia datasets, comparing to Label Power-set, Classifier Chains and Ensembles of Classifier Chains methods. The comparative analysis of the methods performance is done by experiments conducted on image, musical, audio and text datasets.

The experiments have shown that Labels Chain technique outperforms convincingly other methods, in terms of Classification Accuracy for 9 out of 12 datasets. In many cases the differences between the measure values are significant. Also loss functions indicated LC as the best techniques for the majority of the datasets. In the case of the text datasets, LC technique showed the best performance taking into account all the evaluation measures. Such promising results entitle for further investigations of LC techniques, especially for text datasets. Future research should consist in experiments on datasets of different number of attributes and sizes.

References

1. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recogn. **37**(9), 1757–1771 (2004)
2. Briggs, F. et al.: The 9th annual MLSP competition: new methods for acoustic classification of multiple simultaneous bird species in a noisy environment. In: IEEE International Workshop on Machine Learning for Signal Processing, pp. 1–8 (2013)
3. Dembczyński, K., Cheng, W., Hullermeier, E.: Bayes optimal multilabel classification via probabilistic classifier chains. In: International Conference on Machine Learning, pp. 1609–1614 (2010)
4. Glinka, K., Zakrzewska, D.: Effective multi-label classification method for multidimensional datasets. In: Andreasen, T. et al. (eds.) Flexible Query Answering Systems 2015. Advances in Intelligent Systems and Computing 400, pp. 127–138. Springer International Publishing, Switzerland (2016)
5. Gonçalves, E.C., Plastino, A., Freitas, A.A.: A genetic algorithm for optimizing the label ordering in multi-label classifier chains. In: EEE 25th International Conference on Tools with Artificial Intelligence (ICTAI 2013), pp. 469–476. Herndon (2013)
6. Hand, D.J., Yu, K.: Idiot's Bayes: not so stupid after all? Int. Stat. Rev. **69**(3), 385–398 (2001)
7. Huang, J. et al.: Group sensitive classifier chains for multi-label classification. In: 2015 IEEE International Conference on Multimedia and Expo, pp. 1–6 (2015)
8. Kajdanowicz, T., Kazienko, P.: Multi-label classification using error correcting output codes. Int. J. Appl. Math. Comput. Sci. **22**(4), 829–840 (2012)
9. Katakis, I., Tsoumakas, G., Vlahavas, I.: Multilabel text classification for automated tag suggestion. In: Proceedings of the ECML/PKDD 2008 Discovery Challenge. Belgium (2008)
10. Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, D.: An extensive experimental comparison of methods for multi-label learning. Pattern Recogn. **45**(9), 3084–3104 (2012)
11. Mulan: A Java Library. http://mulan.sourceforge.net/datasets-mlc.html
12. Pestian, J.P. et al.: A shared task involving multi-label classification of clinical free text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational and Clinical Language Processing, pp. 97–104. Stroudsburg, PA, USA (2007)

13. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) Machine Learning and Knowledge Discovery in Databases. LNCS 5782, pp. 254–269. Springer (2009)
14. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Mach. Learn. **85**(3), 335–359 (2011)
15. Schapire, R.E., Singer, Y.: BoosTexter: a boosting-based system for text categorization. Mach. Learn. **39**(2/3), 135–168 (2000)
16. Trajdos, P., Kurzynski, P.: An extension of multi-label binary relevance models based on randomized reference classifier and local fuzzy confusion matrix. In: Jackowski, K. et al. (eds.) Intelligent Data Engineering and Automated Learning—IDEAL 2015, LNSC 9375, pp. 69–76. Springer International Publishing (2016)
17. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music into emotions. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008), pp. 325–330. Philadelphia, PA, USA (2008)
18. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 667–685. Springer, US, Boston, MA (2010)
19. Weka Software. http://www.cs.waikato.ac.nz/ml/weka/index.html
20. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, USA (2011)
21. Ye, Ch., Wu, J., Sheng, V.S., Zhao, P., Cui, Z.: Multi-label active learning with label correlation for image classification. In: 2015 IEEE International Conference of Image Processing (ICIP), pp. 3437–3441 (2015)
22. Zhang, B., Wang, Y., Chen, F.: Multilabel image classification via high-order level correlation driven active learning. IEEE Trans. Image Process. **23**(3), 1430–1441 (2014)
23. Zhu, S., Ji, X., Xu, W., Gong, Y.: Multi-labelled classification using maximum entropy method. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 274–281. New York, USA (2005)

Implementing Statistical Machine Translation into Mobile Augmented Reality Systems

Krzysztof Wołk, Agnieszka Wołk and Krzysztof Marasek

Abstract A statistical machine translation (SMT) capability would be very useful in augmented reality (AR) systems. For example, translating and displaying text in a smart phone camera image would be useful to a traveler needing to read signs and restaurant menus, or reading medical documents when a medical problem arises when visiting a foreign country. Such system would also be useful for foreign students to translate lectures in real time on their mobile devices. However, SMT quality has been neglected in AR systems research, which has focused on other aspects, such as image processing, optical character recognition (OCR), distributed architectures, and user interaction. In addition, general-purpose translation services, such as Google Translate, used in some AR systems are not well-tuned to produce high-quality translations in specific domains and are Internet connection dependent. This research devised SMT methods and evaluated their performance for potential use in AR systems. We give particular attention to domain-adapted SMT systems, in which an SMT capability is tuned to a particular domain of text to increase translation quality. We focus on translation between the Polish and English languages, which presents a number of challenges due to fundamental linguistic differences. However, the SMT systems used are readily extensible to other language pairs. SMT techniques are applied to two domains in translation experiments: European Medicines Agency (EMEA) medical leaflets and the Technology, Entertainment, Design (TED) lectures. In addition, field experiments are conducted on random samples of Polish text found in city signs, posters, restaurant menus, lectures on biology and computer science, and medical leaflets. Texts from these

K. Wołk (✉) · A. Wołk (✉) · K. Marasek
Polish-Japanese Academy of Information Technology, ul. Koszykowa 86,
02-008 Warsaw, Poland
e-mail: kwolk@pja.edu.pl; kwolk@pjwstk.edu.pl

A. Wołk
e-mail: awolk@pja.edu.pl

K. Marasek
e-mail: kmarasek@pja.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_6

domains are translated by a number of SMT system variants, and the systems' performance is evaluated by standard translation performance metrics and compared. The results appear very promising and encourage future applications of SMT to AR systems.

Keywords Statistical machine translation · Augmented reality translation · NLP

1 Introduction

An augmented reality (AR) system combines actual objects with computer-generated information to enrich the interaction of a user with the real world [1]. When many people think of AR, they think of the historical head-mounted display (HMD) worn by users. However, AR is not limited to display technologies such as the HMD. For example, video from a smart phone can be enhanced with information that is annotated on the video image. The ubiquity and affordability of the smart phone has led many researchers to investigate different AR applications that leverage smart phone capabilities.

We find written text all around us, but in many different languages [2]. A user of an AR system may not know the language of such text. One such application is translating foreign language text in images taken by a smart phone. There are many potential uses for such a capability. For example, it would be useful for a traveler to be able to read signs, posters, and restaurant menus when visiting a foreign country. A traveler may also have need of medical attention in another country. Having the ability to read handheld medical documents, translated to their own language, using a smart phone would be a tremendous help to these individuals. In addition, a similar capability for a physician attending to a traveler to read medical documents written in the traveler's language would be very useful [3, 4].

There are several aspects to translating text from a smart phone video image. Image processing techniques must be used to identify text and extract it from the image. This may require various enhancements of the image. Next, optical character recognition (OCR) must be applied to the text in an image to distinguish the characters in a particular language, producing a string of text in that language. Lastly, the text can then be automatically translated from the source language to another target language.

Machine translation (MT), translation by a computer in an automated fashion, is a very challenging step for an AR system. Modern MT systems are not based on the rules that specific languages follow. Statistical machine translation (SMT) systems apply statistical analysis to a very large volume of training data for a pair of languages. These statistics, used to tune an SMT system, can then be used to translate new texts from one of the languages to another. If an SMT system is trained for a particular domain of texts (e.g., road signs or medical texts), then it can produce better results on new texts from that domain [5].

The SMT aspects of an AR system are critical. If the end objective is to augment a display with translated text, then translation is obviously important.

A mistranslated road sign could send a traveler in the opposite direction, even resulting in a safety problem in a foreign country. Misunderstandings between a physician and a patient speaking different languages about medical documents could have more serious consequences, such as a misdiagnosis that results in incorrect treatment and corresponding health consequences.

Translation of medical leaflets for patients is another application. Students may find it useful to obtain translated lectures on subjects such as biology and computer science.

The quality of machine translation has greatly improved over time. Several general-purpose translation services, such as Google Translate [6], are currently available on the web. However, they are intended to translate text from a wide variety of domains, and are therefore neither perfect nor well-tuned to produce high quality translations in specific domains.

While our research is extensible to other language pairs, we focus on the challenge of Polish and English. Polish is classified in the Lechitic language group, a branch of the West Slavic language family that includes Czech, Polish, and Slovak. The Polish language uses both animate and inanimate nouns, seven cases, and three genders. The agreement of adjectives with nouns is required in number, case, and gender. Polish has complex rules for applying these elements, and many words are borrowed from other languages. Its vocabulary is enormous. All this together makes the language a very significant challenge for SMT. Likewise, English is known to present considerable translation challenges, and the syntaxes of the two languages are considerably different. The typical gross mismatch in vocabulary sizes between Polish and English also complicates the SMT process [7].

This paper will first review the state of the art in machine translation for AR systems. Next, we will describe our research approach, preparation of the language data, and methods of SMT evaluation. Experiments will be described and their results presented. Lastly, we will discuss the results and draw conclusions.

2 Review of State of the Art

An AR system that enables a user wearing a head-mounted display to interact with a handheld document is presented in [8]. Using their system, a user can select words or very short phrases in the document to access a dictionary. This mechanism could be used to select words or short phrases for translation. However, the focus of their work is on merging a document tracking method with a finger pointing method to enable interaction with the handheld document (via the imager worn by the user). Language translation is not the focus of that research, nor is it demonstrated along with their system. In addition, this approach could only translate single words or short phrases that are pre-segmented by the user. The user must select the text for processing.

The authors of [2] describe a system that extracts and translates a word, once designated by a user in an area of an image. Much of its processing runs on a smart

phone's processor. However, the focus of their research is image processing. The system requires an Internet connection and uses Google Translate to perform the text translation."TextGrabber + Translator" [9] is an Android phone-based application that extracts and translates from a variety of printed sources by using a phone's camera. This application also uses Google Translate. Researchers in [10] describe a Microsoft "Snap and Translate" mobile application that extracts text (designated by a user), performs OCR, and translates between English and Chinese. This prototype system uses a client and cloud architecture, with Bing Translator as the translation service. The authors do not present translation accuracy results.

The Word Lens translation application for Apple iPhone images is described in [11, 12]. Accuracy is mentioned as a challenge. Word Lens only translates words and short phrases, not entire texts. EnWo, discussed in [4], is a similar Android application that translates among several languages. Both applications are commercial applications, not open source, which limits their availability. A system that translates text captured by a mobile phone's camera is briefly described in [13]. This system displays overlays of translated words on the phone's camera. However, neither the translation approach nor its accuracy was reported. The focus of this effort appears to be techniques for OCR and augmenting displays for the user.

The authors of [14] describe a prototype Android phone-based system that detects and translates text in the Bangla (a.k.a. Bengali) language. This system uses the Tesseract engine for performing OCR and Google Translate for translation into English. The authors of [1] describe a translator application, iTranslatAR, that translates text in images or manually-entered text on the iOS platform to augment the missing hearing sense and for other applications. This prototype application uses Tesseract OCR and Google Translate as well.

The authors of [15] describe a system that uses genetic algorithms to recognize text in smart phone camera images, such as road signs, and makes it available for translation from English to Tamil through online translation services. However, this research is focused on image processing and reducing the bandwidth required for images. The quality of the translation itself is limited to online translation services. It is also unclear whether or not text at different distances can be processed by this prototype system. In addition, the accuracy of the resulting translations was not evaluated.

A head-mounted text translation system is described in [16]. This research focuses on eye gaze gestures in a head-mounted display and an innovative OCR method. These researchers implemented an embedded Japanese-to-English translation function that is based on common Japanese words translated using Microsoft's Bing translator. However, the accuracy of the translation was not assessed.

A text recognition and translation system hosted on a smart phone and web-based server is discussed in [3]. The user of this system specifies the region of an image, obtained by the phone's camera that contains text. The image is processed to limit required bandwidth and transmitted to the server, where text recognition and translation is performed. The prototype system, which focused on English and simplified Chinese as two very challenging languages, uses the open-source Tesseract OCR engine and Google Translate for language translation.

Due to the challenge of recognizing text, a user can correct several aspects of character classification to improve accuracy. Evaluation showed poor phrase match accuracy (under 50 % in all cases).

We observe from this literature that machine translation performance has not been the focus of research for augmented reality systems. Concluding from described examples, most research efforts have emphasized image processing techniques, OCR, mobile-network architecture, or user interaction approaches. This is further indicated by the frequent lack of reporting on translation accuracy results from these studies.

In addition, previous studies have typically used online translation services, scaled-down versions of them, or proprietary software in the language translation portion of their systems. Some of the translation systems used are commercial and proprietary, not open source and not readily available. Even the online translation systems are often not free for AR applications. For example, Google charges based on usage for their Google Translate API as part of the Google Cloud Platform [6]. An online translation service, such as Google, also requires a constant Internet connection for an AR system to perform translation that introduces lags.

The lack of focus on domain-adapted machine translation systems for augmented reality is particularly noteworthy. Such general purpose translation systems collect bilingual data from wherever it is available to gather as much data as possible. However, the domain (e.g., medical texts or road signs) to which the resulting translation system is applied may be drastically different from that general data. Our approach is to use in-domain adaptation so that a translation system can be tuned for a specific narrow domain. Research has shown that such adaptation can improve translation accuracy in a specific domain [5, 17]. In-domain adaptation promises to provide similar benefits to AR applications because it also reduces computation costs and is some cases makes it possible to run SMT engine on the handheld device.

3 Approach

This research effort developed a tool that recognizes text in a camera image, translates it between two languages, and then augments the camera image (real-time on device screen) with the translated text. As discussed in the review of the state-of-the-art, such tools are already on the market, but most of them translate only single words like city signs or work in a less then fully-automated fashion. This work is innovative because it works in real time and translates an entire text, paragraphs, or pages. In this manner, the translation's context is maintained, resulting in an improved translation quality. For example, someone might go to a restaurant and view a menu that would be automatically translated on their mobile device.

Translation systems work best if their domain is limited. General purpose translation systems, on the other hand, produce lower-quality translations in very

specific domains. For example, a general purpose system such as Google Translation would not do well in translating medical leaflets, since medicine is a very specific and specialized domain.

The approach adopted in this research uses a unique SMT systems trained for specific domains. This should produce higher-quality translations then a general purpose translation system. The domains selected for SMT adaptation are: medical leaflets, restaurant menus, news, and article or lecture translation.

The tool developed during this research translates all the text it recognizes in the camera image. Most tools from the reviewed literature translate single words or small, selected phrases. This results in those approaches loose the context of the translation. Loss of that context can only lead to lower quality. In addition, our approach is easy to implement and also based on open source tools, not restricted in use, as are commercial solutions. It is also extensible to any language pair, not just limited to the languages used in our experiments.

4 Data Preparation

This research used Polish and English language data from two primary and distinct parallel corpora: the European Medicines Agency (EMEA) and the 2013 Technology, Entertainment, Design (TED) lectures. The EMEA data is composed of approximately 1,500 biomedical documents in Portable Document Format (PDF)—approximately 80 MB of data in UTF-8 format—that relate to medicinal products. These documents are translated into the 22 official European Union languages. Separation of these texts into sentences yielded 1,044,764 sentences, which were constructed from 11.67 M untokenized words. The disproportionate Polish-English vocabulary in the EMEA data is composed of 148,170 and 109,326 unique Polish and English words, respectively.

Errors in the Polish EMEA data first had to be repaired. Additional processing of the Polish data, which used the Moses SMT processing toolkit [18], was required prior to its use in training the SMT model. This toolkit includes tools for creating, training, tuning, and testing SMT models. Removal of long sentences (defined as 80 tokens) was part of this preprocessing. Preparation of the English data was much simpler than that for Polish, and there were fewer errors in the data. However, the removal of strange UTF-8 symbols and foreign words was required.

The TED lecture data, totaling approximately 17 MB, included approximately 2.5 M untokenized Polish words, which were separated into sentences and aligned in language pairs. Similar to [19], the data was preprocessed using both automated cleaning and manual changes to correct problems, including spelling errors, nesting problems (described in [19]), and repeated text fragments that caused issues in text parallelism. The authors describe the tool used to correct many of these problems in [20]. Preparation of the English data was, once again, much simpler than that of the Polish data. The main focus was cleaning of the data to eliminate errors.

As in the EMEA data, this processing of TED lecture data resulted in a disproportionate vocabulary of 59,296 and 144,115 unique English and Polish words, respectively. However, the domain of the TED lectures was much less specific and more wide-ranging then that of the EMEA data.

5 Methods of Evaluation

Since human evaluations of the quality of translated output are very expensive and time-consuming, automated SMT quality metrics must be calculated for any significant amount of data. Since SMT systems rely on processing a lot of data, automated metrics are typically required.

There has been much research on automated metrics. From this research, four primary, language-independent measures are preferred in the scientific community: Bilingual Evaluation Understudy (BLEU), the U.S. National Institute of Standards & Technology (NIST) metric, Metric for Evaluation of Translation with Explicit Ordering (METEOR), and Translation Error Rate (TER). These measures are used to compare a reference translation of presumably high quality to the translation output by an SMT system [21].

BLEU is a widely-used metric that is very easy and quick to calculate. This metric uses weighted averages to match varying-length phrases from the translations being compared, measuring the overlap in words and giving higher scores to matching phrases (word sequences). The position of words and phrases in the text is not considered by this metric. To avoid a bias that might artificially inflate the scores of an SMT system through the high usage of high-confidence words, the total word count is considered in this metric. In addition, a penalty is applied for too much brevity in a translation. BLEU uses the geometric mean of sentence-by-sentence scores to evaluate translation performance on a complete corpus [21, 22].

The NIST metric was intended to improve BLEU, in part to reward more the correct translation of infrequently-used words. This is accomplished through a stronger weighting of rare words and the use of the arithmetic mean of the matches to summarize scores for a text. The NIST metric is also less sensitive to phrase length through modification of the BLEU brevity penalty. Improvements of this metric over BLEU have been demonstrated in [22].

Innovations in the METEOR metric include the consideration of higher-order n-grams to reward word order matches in translations, as well as use of the arithmetic mean (like NIST) for document scores. In addition, METEOR modifies the BLEU brevity penalty and emphasizes the use of multiple reference translations (instead of attempting to match only one reference translation) [22, 23].

Unlike the previously-mentioned SMT metrics, a TER score is better if it is lower. TER is a more recently developed translation metric and takes a very different approach. It measures the number of word insertions and deletions, word and phrase order changes, and substitutions of words a human translator would need to

make for SMT output to match a reference translation. The objective of this metric is to address the meaning and fluency of a translation [24].

To provide a comprehensive evaluation of our SMT approach for an AR system, all four of these metrics were used in the evaluation of our experimental results.

6 Experiments

Experiments were conducted to evaluate different SMT systems using various data. In general, each corpus was tokenized, cleaned, factorized, converted to lowercase, and split. A final cleaning was performed. For each experiment, the SMT system was trained, a language model was applied, and tuning was performed. The experiments were then performed. For OCR purposes we used well known from good quality Tesseract engine [25], by this we also evaluated an impact of OCR mistakes on translation. It also must be noted that the OCR system was not adapted to any specific types of images or texts which would most probably improve its quality. It was not done because it was not goal of this research.

The Moses toolkit, its Experiment Management System, and the KenLM language modeling library [26] were used to train the SMT systems and conduct the experiments. Training included use of a 5-gram language model based on Kneser-Ney discounting. SyMGIZA++ [27], a multi-threaded and symmetrized version of the popular GIZA++ tool [18], was used to apply a symmetrizing method to ensure appropriate word alignment. Two-way alignments were obtained and structured, leaving the alignment points that appear in both alignments. Additional points of alignment that appear in their union were then combined. The points of alignment between the words in which at least one was unaligned were then combined (grow-diag-final). This approach facilitates an optimal determination of points of alignment [28]. The language model was binarized by applying the KenLM tool [26].

Finally, the approach described by Durrani et al. in [29] was applied to address out-of-vocabulary (OOV) words found in our test data. Despite the use of large amounts of language data, OOV words, such as technical terms and named entities, pose a challenge for SMT systems. The method adopted here from [29] is a completely unsupervised and language-independent approach to OOV words.

Experiments were performed using traditional SMT systems on EMEA medical leaflet data and then on TED lecture data, translating Polish to English. These experiments were repeated for texts taken randomly from the corpora as well as on the same texts recognized by the OCR system. Data was obtained through the process of random selection of 1,000 sentences and their removal from the corpora. The BLEU, NIST, TER, and METEOR metrics were used to evaluate the experimental results.

Tables 1 and 2 provide the experimental results for translation. The EMEA abbreviation stands for translation of medical leaflets, TED for discussed earlier TED corpus and FIELD for field experiments that were performed on random

Table 1 Polish-to-English translation

System	BLEU	NIST	METEOR	TER
EMEA	76.34	10.99	85.17	24.77
EMEA-OCR	73.58	10.77	79.04	27.62
TED	28.62	6.71	58.48	57.10
TED-OCR	27.41	6.51	56.23	60.02
FIELD	37.41	7.89	63.76	45.31
FIELD-OCR	35.39	7.63	61.23	48.14

Table 2 English-to-Polish translation

System	BLEU	NIST	METEOR	TER
EMEA	73.32	10.48	81.72	27.05
EMEA-OCR	69.15	9.89	75.91	29.45
TED	26.61	5.99	48.44	59.94
TED-OCR	23.74	5.73	46.18	61.23
FIELD	33.76	7.14	61.68	49.56
FIELD-OCR	32.11	6.99	59.17	52.05

samples of Polish city signs, posters, restaurant menus, lectures on biology and computer science, and medicine boxes. The field SMT system was trained on both TED and EMEA corpora. The images were taken in different situations, lightning and devices. Experiments that have additional "–OCR" suffix in their name were first processed by OCR engine and secondly translated.

Translation phrase tables often grow enormously large, because they contain a lot of noisy data and store many very unlikely hypotheses. This is problematic when a real-time translation system that requires loading into memory is required especially on mobile devices with limited resources. We decided to try a method of pruning those tables introduced by Johnson et al. [30] and our method based on dictionary.

In Table 3, experiments are shown (Absolute N—keep sentence if at least N words from dictionary appear in it, relative N—keep a sentence if at least N% of sentence is built by from dictionary, pruning N—keep only N most probable translations). We conducted those experiments on most important translation engine which is the field test starting with OCR from Polish to English language.

Contrary to the referenced publications, the empirical experiments showed that pruning (for PL-EN) decreases translation quality significantly. Even though phrase and reordering tables are much smaller, the decrease in quality is questionable. What is interesting is that a substantial factor of quality loss to final phrase table size was obtained by combining pre-filtering and pruning of a phrase table in Experiment 12.

Table 3 Pruning results

Optimizer	BLEU	Phrase table (GB)	Reordering table (GB)
None	35.39	6.4	2.3
Absolute 3	26.95	1.1	0.4
Absolute 2	30.53	2.5	0.9
Absolute 1	32.07	4.9	1.7
Relative 2.5	30.82	3.1	1.1
Relative 5	26.35	1.1	0.4
Relative 7.5	17.68	0.3	0.1
Pruning 30	32.36	1.9	0.7
Pruning 60	32.12	2.0	0.7
Pruning 90	32.11	2.0	0.75
Pruning 20	32.44	2.1	0.75
Absolute 1 + Pruning 20	30.29	0.85	0.3

7 Discussion and Conclusions

AR systems would greatly benefit from the application of state-of-the-art SMT systems. For example, translation and display of text in a smart phone image would enable a traveler to read medical documents, signs, and restaurant menus in a foreign language. In this paper, we have reviewed the state of the art in AR systems, described our SMT research approach—including the important preparation of language data and models, as well as evaluation methods—and presented experimental results for several different variants of Polish-English SMT systems.

Clearly, machine translation performance has not been the focus of AR research. The literature shows an emphasis on image processing techniques, OCR, mobile-network architecture, and user interaction approaches. As a result, previous AR research has generally used general-purpose translation services, which are not tuned for specific text domains and do not produce the highest quality translations. In addition, AR research reported in the literature frequently lack reporting of translation accuracy results.

The focus of this research was machine translation between the Polish and English languages. Both languages present significant challenges due to vast differences in syntax, language rules, and vocabulary. Experiments were conducted in Polish-to-English and English-to-Polish direction on the EMEA and TED data. In addition, field experiments were performed on a variety of random samples in Polish. Standard machine translation evaluation methods were used to evaluate the results, which show great promise for the eventual SMT use in AR systems.

BLEU scores lower than 15 mean that the machine translation engine is unable to provide satisfactory quality, as reported by Lavie [31] and a commercial software manufacturer [32]. A high level of post-editing will be required to finalize output translations and reach publishable quality. A system scores greater than 30 %

means that translations should be understandable without problems. Scores over 50 reflect good and fluent translations.

Overall, the EMEA and TED experimental results appear adequate within the limits of the specific text domain. SMT is particularly useful for the complex Polish language. However, there is clearly room for future improvement for critical applications. English-to-Polish translation proved more challenging, as expected. The results from the OCR experiment generally show translation performance only a bit lower then plain SMT. Most likely reason is that statistical translation methods are not vulnerable to small mistakes like spelling etc. All the results were very encouraging, but more work remains to further optimize SMT translation quality. The results also reinforce the importance of phrase table pruning so it can be small enough to fit into mobile device memory. In our experiments we were able to obtain such threshold by reducing quality by about 5 BLEU points.

There are many advantages to the SMT approach we took in our research. First, an advanced SMT system was adapted for particular text domains. Such a domain-adapted system produces higher-quality translations then general-purpose translation services, such as Google Translate, in specific domains. In addition, our approach is not restricted by commercial license and is easy to implement.

The proposed SMT approach, unlike other AR systems, translates all the text it sees without need to transfer data over Internet. This promotes automation and flexibility in an AR system user, since they do not have to identify specific text for translation and are not limited to simple words or phrases. Translating larger texts also leverages the value of context, which increases translation quality. Lastly, our approach is easily extensible to any language pair.

Acknowledgments This research was supported by Polish-Japanese Academy of Information Technology statutory resources (ST/MUL/2016) and resources for young researchers.

References

1. Carmigniani, J.: Augmented Reality Methods and Algorithms for Hearing Augmentation. Florida Atlantic University (2011)
2. Fragoso, V., et al.: Translatar: a mobile augmented reality translator. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 497–502. IEEE (2011)
3. Gschwandtner, M., Kritz, M., Boyer, C.: Requirements of the health professional research. Technical Report (2011)
4. NeWo LLC.: EnWo—English cam translator. https://play.google.com/store/apps/details?id=com.newo.enwo. Accessed 12 Jan 2014
5. Cui, L., et al.: Multi-domain adaptation for SMT using multi-task learning. In: EMNLP, pp. 1055–1065 (2013)
6. Google.: Translate API. https://cloud.google.com/translate/v2/pricing. Accessed 6 April 2015
7. Wołk, K., Marasek, K.: A sentence meaning based alignment method for parallel text corpora preparation. New Perspectives in Information Systems and Technologies, vol. 1, pp. 229–237. Springer International Publishing (2014)

8. Martedi, S., Uchiyama, H., Saito, H.: Clickable augmented documents. In: 2010 IEEE International Workshop on Multimedia Signal Processing (MMSP), pp. 162–66. IEEE (2010)
9. Abbyy.: TextGrabber + Translator. http://abbyy-textgrabber.android.informer.com/. Accessed 16 Feb 2015
10. Du, J., et al.: Snap and translate using windows phone. In: 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 809–813. IEEE (2011)
11. The Economist. Word Lens: This changes everything. http://www.economist.com/blogs/gulliver/2010/12/instant_translation. Accessed 18 Dec 2010
12. DATAMARK, Inc.: OCR as a real-time language translator. https://www.datamark.net/blog/ocr-as-a-real-time-language-translator. Accessed 30 July 2012
13. Khan, T., et al.: Augmented reality based word translator. Int. J. Innov. Res. Comput. Sci. Technol. (IJIRCST) **2**(2) (2014)
14. Mahbub-Uz-Zaman, S., Islam, T.: Application of augmented reality: mobile camera based bangla text detection and translation. 2012. PhD Thesis. BRAC University (2012)
15. Emmanuel, A.S., Nithyanandam, S.: An optimal text recognition and translation system for smart phones using genetic programming and cloud. Int. J. Eng. Sci. Innov. Technol. (IJESIT) **3**(2), 437–443 (2014)
16. Toyama, T., et al.: A mixed reality head-mounted text translation system using eye gaze input. In: Proceedings of the 19th International Conference on Intelligent User Interfaces, pp. 329–334. ACM (2014)
17. Koehn, P., Schroeder, J.: Experiments in domain adaptation for statistical machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics, pp. 224–227 (2007)
18. Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, pp. 177–180 (2007)
19. Wołk, K., Marasek, K.: Polish–English speech statistical machine translation systems for the IWSLT 2013. In: Proceedings of the 10th International Workshop on Spoken Language Translation, Heidelberg, Germany, pp. 113–119 (2013)
20. Wołk, K., Marasek, K.: A sentence meaning based alignment method for parallel text corpora preparation. New Perspectives in Information Systems and Technologies, vol. 1, pp. 229–237. Springer International Publishing (2014)
21. Koehn, P. What is a better translation? Reflections on six years of running evaluation campaigns. Tralogy (2011)
22. Wołk, K., Marasek, K.: Enhanced bilingual evaluation understudy. Lect. Notes Inf. Theory **2** (2) (2014)
23. Banerjee, S., Lavie, A.: Meteor: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
24. Snover, M., et al.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231 (2006)
25. Smith, R.: An overview of the tesseract OCR engine. ICDAR, pp. 629–633. IEEE (2007)
26. Heafield, K.: KenLM: faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, pp. 187–197 (2011)
27. Information Systems Laboratory, Adam Mickiewicz University.: SyMGIZA++. http://psi.amu.edu.pl/en/index.php?title=SyMGIZA. Accessed 27 April 2011
28. Gao, Q., Vogel, S.: Parallel implementations of word alignment tool. Software Engineering, Testing, and Quality Assurance for Natural Language Processing. Association for Computational Linguistics, pp. 49–57 (2008)
29. Durrani, N., et al.: Integrating an unsupervised transliteration model into statistical machine translation. EACL **2014**, 148 (2014)

30. Johnson, J.H., et al.: Improving translation quality by discarding most of the phrase table. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (2007)
31. Lavie, A.: Evaluating the output of machine translation systems. AMTA Tutorial (2010)
32. KantanMT—a sophisticated and powerful machine translation solution in an easy-to-use package. http://www.kantanmt.com

The Use of the Universal Quality Index for User Recognition Based on Fingerprint Analysis

Jakub Peksinski, Grzegorz Mikolajczak and Janusz Kowalski

Abstract In the article, the authors presented the possibilities of using the Universal Image Quality Index (Q)—a popular measure for evaluation of digital image quality in order to identify users based on analysing their fingerprints with the use of a reference image. The applied quality measure is used both for analysing of fingerprints as well as in the process of synchronisation preceding the analysis.

Keywords Digital image · Data matching · Fingerprint analysis

1 Introduction

Since time immemorial, people have tried to create methods which would enable identification of an individual in an explicit manner that does not raise any doubts, based on certain individual and unique characteristic features.

Such identification may be carried out with the use of various techniques. The identification techniques currently in use may be divided into three basic groups [1, 2]—identification techniques using the knowledge of an individual for their operation, techniques based on electronic identifiers, identification techniques using a comparison of unique features possessed by an individual for their operation.

The first group includes all identification techniques which use a string of alphanumeric characters such as e.g. passwords, PIN codes, etc.

J. Peksinski (✉) · G. Mikolajczak
Faculty of Electrical Engineering, West Pomeranian University of Technology, Szczecin, Poland
e-mail: jpeksinski@zut.edu.pl

G. Mikolajczak
e-mail: grzegorz.mikolajczak@zut.edu.pl

J. Kowalski
Faculty of Medicine, Pomeranian Medical University, Szczecin, Poland
e-mail: janus@pum.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_7

The second group includes methods using electronic devices such as e.g. magnetic cards, etc.

The third group includes identification techniques which use the analysis of biological features of an individual, i.e. biometric parameters.

The operation of biometric techniques is based on the analysis of unique features of each individual. Basic biometric techniques included examination of the following [3]: fingerprints, hand geometry, face geometry, retina, iris, DNA.

Currently, a very frequently used method of biometric analysis is comparison of fingerprints, which every human being has. Each human being has a unique pattern of epidermal ridges and the probability of another person having an identical pattern is 1:64 billion. The pattern consists of specific unique elements such as [4]: minutiae, the arrangement of epidermal ridges, the location of pores, the shape of pores, the shape of epidermal ridge edges.

The main task for the identification system is proper read-out of these characteristic features of an individual—processing them into a digital form and then comparing the obtained digital representation with the biometric model of a given individual stored in the data base.

While creating a biometric system, a decision must be made which feature will be used as the biometric identifier. The selected feature must fulfil the following conditions [5, 6]—universality—which each human being has; uniqueness—any two people can be explicitly distinguished with the use of the selected feature; changelessness—the feature does not change with the flow of time; easy obtain ability—the feature can easily be measured and transformed into a digital form.

The aim of the article is for the authors to develop a method of proper user identification based on fingerprint analysis. In order to achieve it, the authors of the article used popular methods and algorithms commonly applied in digital analysis and image processing.

The field of science dealing with the issues of digital image processing includes a whole range of issues beginning with the process of digital image recording, through its processing, analysing and storing, ending with its projection on a monitor or telephone screen, etc. [7–9]. The broad group of issues which includes digital image processing within its scope also covers the issues connected with: digital image quality assessment, geometric transformations on the image which are included in the group of so called point transformations.

The authors of the article focused on two of the above mentioned issues because the following where used to achieve the expected aim, which is proper identification of an individual based on fingerprint analysis:

1. A popular measure used to assess digital image quality—Universal Quality Image Index [Q] [9].

$$Q = \frac{4 \cdot \sigma_{x,y} \cdot \tilde{x} \cdot \tilde{y}}{\left(\sigma_x^2 + \sigma_y^2\right) \cdot (\tilde{x}^2 + \tilde{y}^2)} \tag{1}$$

where: $\tilde{x} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$; $\tilde{y} = \frac{1}{N} \cdot \sum_{i=1}^{N} y_i$; $\sigma_x = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \tilde{x})^2$; $\sigma_y = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (y_i - \tilde{y})^2$; $\sigma_{x,y} = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \tilde{x}) \cdot (y_i - \tilde{y})$

The value of [Q] measure indication depends on three factors: correlation, luminance, contrast.

2. Well-known and commonly used operations performed on digital images including:

- Translation:

$$x_n = x_0 + a \quad y_n = y_0 + b \tag{2}$$

where: x_0, y_0—old image coordinate; x_n, y_n—new image coordinate; *a, b*—translation value;

- Image rotation by a certain angle:

$$\begin{aligned} x_n &= A + (x_1 - A) \cdot \cos(\alpha) - (y_1 - B) \cdot \sin(\alpha) \\ y_n &= B + (x_1 - A) \cdot \sin(\alpha) + (y_1 - B) \cdot \cos(\alpha) \end{aligned} \tag{3}$$

where: x_n, y_n—the value of new pixel location after rotation; x_1, y_1—the value of location before rotation, *A, B*—rotation axis, α—rotation angle

2 The Use of [Q] Indication in Fingerprint Analysis

In order to demonstrate the possibility of using the [Q] measure indication in the process of user identification, Example 1 was used presenting the effectiveness of [Q] measure indication in the process of user identification based on digital image analysis of the fingerprint. All fingerprints were taken with a scanner with CCD matrix being a part of a popular multi-function device—HP DeskJet Ink Advantage with the resolution of 200 dpi.

Example 1 Figure 1 presents a digital representation of an individual's fingerprint. The fingerprint presented in Fig. 1a is a reference model stored in the base. Whereas fingerprint in Fig. 1b is taken on a current basis.

The fingerprints in Fig. 1a, b were compared with the use of [Q] measure indication. The comparison was carried out in two manners:

Method 1—involves the so called overall comparison of two fingerprint representations 1a and 1b. Table 1 presents example results of [Q] measure indication for various users. As it is visible in the results presented in Table 1, the [Q] measure indication confirms its effectiveness in user identification based on a digital image of a fingerprint. The [Q] measure indication for cases where (a) and (b) fingerprints belong to the same person is distinctly high and fits within the limits from **0.863** to

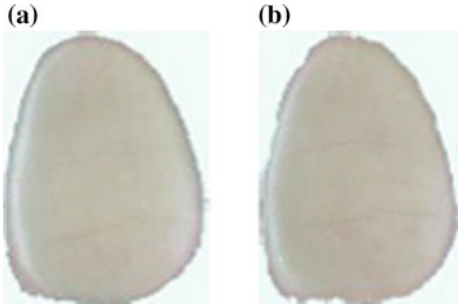


Fig. 1 Digital fingerprint obtained with the use of a scanner with the resolution of 200 dpi

Table 1 The results of Q measure indication for overall comparison

Fingerprints	Fingerprints of the same person	Q indications
Pair 1a and 1b	Yes	0.971
Pair 2a and 2b	Yes	0.863
Pair 3a and 3b	Yes	0.875
Pair 3a and 3b	Yes	0.952
Pair 4a and 4b	Yes	0.958
Pair 5a and 5b	Yes	0.878
Pair 6a and 6b	No	0.221
Pair 7a and 7b	No	0.275
Pair 8a and 8b	No	0.197
Pair 9a and 9b	No	0.211
Pair 10 a and 10b	No	0.131

a—reference models stored in the base; b—images taken

0.981. The [Q] measure indication for cases where (a) and (b) fingerprints belong to two different people is distinctly low and fits within the limits from **0.131** to **0.397**.

Method 2—the so called partial comparison involves comparing of individual pairs of segments with the use of [Q] measure indication. Both representations (a) and (b) were divided into segments 1–9 and 1′–9′ in such a way that each segment from image (a) was assigned precisely the same and located precisely in the same place on the X and Y plane in image (b). In this way, 9 pairs of segments 1-1′, 2-2′,…, 9-9′ were obtained. This is presented in Fig. 2. Example research results are presented in Tables 2 and 3.

Analyzing the results presented in Tables 2 and 3, it is explicitly visible that the Q measure indication for the so called partial method correlates with the results obtained with the overall indication presented in table number 1. It is visible here that the individual indications obtained for segments were different depending on the segment.

Nevertheless, the value of the average indication does not significantly diverge from the overall indication e.g. for:

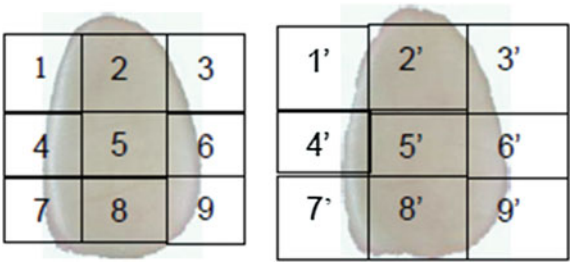


Fig. 2 Digital fingerprint obtained with the use of a scanner with the resolution of 200 dpi—divided into segments

Table 2 The results of Q measure indication for comparing of individual segments for the same person

Fingerprints	Fingerprints of the same person	Q indications
Pair 1a and 1b	Segment 1-1′	0.931
	Segment 2-2′	0.946
	Segment 3-3′	0.924
	Segment 4-4′	0.948
	Segment 5-5′	0.981
	Segment 6-6′	0.979
	Segment 7-7′	0.958
	Segment 8-8′	0.965
	Segment 9-9′	0.957
	Indication average	0.954

Table 3 The results of Q measure indication for comparing of individual segments for two different people

Fingerprints	Fingerprints of the different person	Q indications
Pair 10a and 10b	Segment 1-1′	0.138
	Segment 2-2′	0.159
	Segment 3-3′	0.111
	Segment 4-4′	0.176
	Segment 5-5′	0.211
	Segment 6-6′	0.172
	Segment 7-7′	0.132
	Segment 8-8′	0.104
	Segment 9-9′	0.091
	Indication average	0.143

Image pair 1a and 1b Q indication = 0.971—overall method
Image pair 1a and 1b average indication = 0.954—partial method
Image pair 10a and 1b Q indication = 0.131—overall method
Image pair 1a and 1b average indication = 0.143—partial method

Both methods presented in Example 1 confirmed the effectiveness and the possibility of using the [Q] measure indication in user identification systems based on digital image processing of a fingerprint. Nevertheless, the authors suggest using the second, i.e. partial method. The advantage of this solution is presented in Example 2.

Example 2 The fingerprints are compared as in Example 1. The difference is that fingerprint (b) has a defect (e.g. resulting from scanning or any other reason), whereas the model stored in the base does not have it. The defect is located in segments 5′ and 8′.

In the overall comparison, the following indication of the measure was obtained: **Q = 0.493**. Such indication suggests that fingerprints belong to two different people. This is not true. In the partial comparison, the indications of the measure for individual segments presented in Table 4 were obtained.

Analysing the results presented in Table 4, it is clearly visible that, in the segments covering the defect with their scope (segments 5′ and 8′), the [Q] measure indications were distinctly low, whereas they where high in other segments. Thanks to such a solution and the use of a respectively large number of segments, it is possible to identify an individual with high accuracy even if the person has any defect on the finger that appeared in the process of scanning or much later than the model stored in the base. The above mentioned example confirmed the effectiveness of the [Q] measure. Its comparative property may successfully be used in user identification systems based on fingerprint analysis. The [Q] measure has one significant disadvantage, which seriously influences the value of its indication. The [Q] measure indication depends strongly on translation. The dependence is presented in Example 3.

Example 3 Figure 3 presents digital representation of a fingerprint like in Examples 1 and 2. The difference is that the fingerprint in Fig. 3b is precisely the

Table 4 The results of Q measure indication for comparing of individual segments—a fingerprint with a scar

Fingerprints	Fingerprints of the same person	Q indications
	Segment 1-1′	0.902
	Segment 2-2′	0.915
	Segment 3-3′	0.888
	Segment 4-4′	0.832
	Segment 5-5′	**0.271**
	Segment 6-6′	0.872
	Segment 7-7′	0.889
	Segment 8-8′	**0.404**
	Segment 9-9′	0.891
	Indication average	0.762

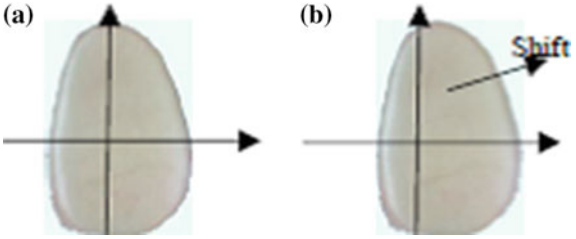


Fig. 3 Digital fingerprint obtained with the use of a scanner with the resolution of 200 dpi—3b shifted on the scanning plane

same as in Fig. 3a only it was shifted by a few pixels. This slight translation has a significant impact on the [Q] measure indication. Table 5 presents the results of the influence of translation between fingerprint representations on the [Q] measure indication for the overall and partial comparison. The test was conducted for fingerprints belonging to the same person. Analysing the results presented in Table 5, it may be concluded that slight translation between the images has a significant impact on the [Q] measure indication value, which seriously influences the final assessment result. As a consequence, it leads to wrong identification.

Therefore, to make the [Q] measure indication effective (enabling explicit identification of an individual) and not burdened with an error resulting from no synchronisation, it is necessary to prepare properly the images representing fingerprints. To this end, an effective method of image synchronisation must be used

Table 5 Q measure indication results for partial and overall comparison with translation

Fingerprints	Translation		Segments	Q in segments	Total Q
	X	Y			
Pair 1a and 1b	−1	0	Segment 1-1′	0.632	0.697
			Segment 2-2′	0.656	
			Segment 3-3′	0.643	
			Segment 4-4′	0.656	
			Segment 5-5′	0.702	
			Segment 6-6′	0.629	
			Segment 7-7′	0.655	
			Segment 8-8′	0.597	
			Segment 9-9′	0.501	
Pair 1a and 1b	−3	0	Segment 1-1′	0.392	0.431
			Segment 2-2′	0.428	
			Segment 3-3′	0.411	
			Segment 4-4′	0.456	
			Segment 5-5′	0.489	
			Segment 6-6′	0.401	
			Segment 7-7′	0.387	
			Segment 8-8′	0.425	
			Segment 9-9′	0.376	

both taking into consideration the translation in the X and Y plane as well as the rotation between the images. In order to achieve this, the authors of the article used typical operations applied on digital images described with Eqs. (2) and (3). They also made use of the negative dependency of the [Q] measure indication on translation described in Example 3, which will be used here to find the value of translation.

3 Finding the Translation Between the Images with the Use of [Q] Measure Indication

In order to make the image synchronization process effective, it is necessary to find translation between the images representing fingerprints. Translation may occur both in the X and Y plane and as a result of rotation between the images.

In order to ensure effectiveness of the synchronization method based on the analysis of the [Q] measure indication, it is necessary to make the following assumptions: fingerprints are taken with the same scanning device with the same resolution; fingerprints are always taken in the same strictly defined area of the scanning plane; fingerprints are always taken in the same position, i.e. there will be no case of e.g. a thumb rotated by e.g. 90° with regard to the scanning plane.

Thanks to such assumptions, it may be considered that: images representing fingerprints will have the same size and resolution; translation between the images will only amount to a few pixels; the rotation angle will only amount to a few degrees. The order of the synchronization algorithm is as follows:

In the first step of the algorithm, it is necessary to check whether the compared images representing fingerprints require synchronization. To this end, the operation of deducting two images from each other is performed. If the difference between them amounts to 0, this will mean that the images are synchronized, otherwise it will be necessary to perform the synchronization process between them.

In the second step it is necessary to find translation in the X and Y plane. To this end, a segment of the image is selected where the [Q] indication is the highest.

The selected segments are shifted against each other in the X and Y plane by one pixel at a time—the [Q] measure indication is calculated in each new location of the segment. Where the [Q] measure indication is the highest (closest to 1), the images will be best synchronized. The process is illustrated in Fig. 4.

Analysing the results presented in Table 6, it may be concluded that the method of finding translation between images based on analysis of [Q] measure indications proposed by the authors of the article is effective. In each case, the pre-defined translation was found in 100 %. Having found the translation, it is possible to synchronise the images in the X and Y plane.

In the third step, the axis and angle of rotation between the examined fingerprints are found.

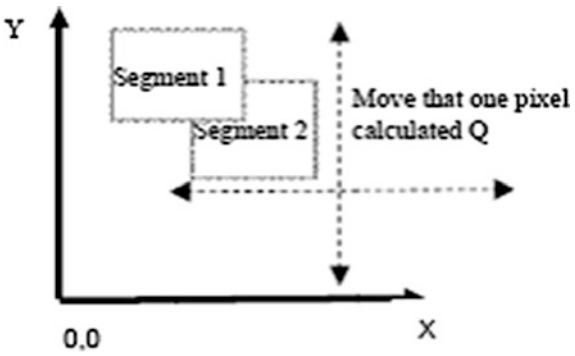


Fig. 4 Finding translation between images

Table 6 The results of finding translation between the images with the use of Q measure

Image pairs	Pre-defined translation		Initial Q indication	Translation found		Final Q indication
	X	Y		X	Y	
Pair 1a-1b	0	1	0.876	0	1	0.973
Pair 2a-2b	0	−2	0.764	0	−2	0.982
Pair 3a-3b	−1	1	0.721	−1	1	0.979
Pair 4a-4b	−2	2	0.699	−2	2	0.978
Pair 5a-5b	−1	−3	0.775	−1	−3	0.987
Pair 6a-6b	1	1	0.923	1	1	0.989
Pair 7a-7b	1	2	0.761	1	2	0.992
Pair 8a-8b	2	1	0.705	2	1	0.903
Pair 9a-9b	3	2	0.651	3	2	0.931
Pair 10a-10b	−2	−2	0.678	−2	−2	0.957

Thanks to the assumptions made, it may be considered that the translation found between the images in direction X and Y located in the segment where the [Q] measure indication is the highest, is the sought axis of rotation between the images. Having determined the axis of rotation, the rotation angle must be found. To this end, the edge of the examined segment is taken that is the furthest away from the rotation axis found (O). In this way, a section of length (a) is obtained with its beginning in the axis of rotation and the end in point X. Next, a mask is selected with dimensions N x N pixels (selected experimentally for a given type of scanner). The mask is moved along the horizontal edge by one pixel at a time beginning from point X. The [Q] indication is calculated in each location of the mask. Point Y is marked where the [Q] indication reaches its maximum value. In this way, a section of length (b) is obtained beginning in point X and ending in point Y.

The process is illustrated in Fig. 5. In this way, a right-angled triangle is obtained with edge length equal to the length of sections a and b found.

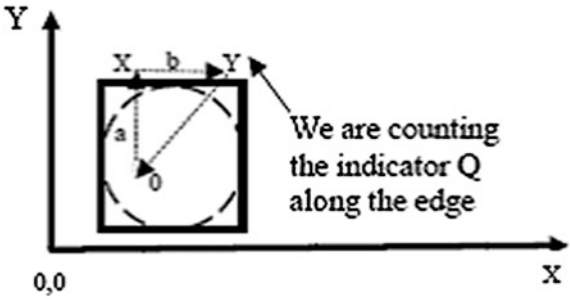


Fig. 5 Finding the rotation angle

The α angle is calculated using the trigonometric dependencies, in accordance with the formula:

$$\alpha = \arctan\left(\frac{b}{a}\right) \tag{4}$$

The effectiveness of finding the rotation angle is confirmed by the experimental test results, examples of which were presented in Table 7—where, as it can be seen, the rotation angle was found each time with 100 % accuracy. For the experimental test, the rotation angle was artificially induced. Having found the axis and angle of rotation, it is possible to proceed to synchronisation of the two images using the dependency described with Eq. (3).

Table 7 Example results of the experimental tests of finding the rotation axis and angle

Image pairs	Axis found		Pre-defined rotation angle	Rotation angle found
	X	Y		
Pair 1a-1b	1.421	1.519	3	3
Pair 2a-2b	1024	1.201	2	2
Pair 3a-3b	2221	2457	4	4
Pair 4a-4b	1339	1567	3	3
Pair 5a-5b	1400	1500	4	4
Pair 6a-6b	1297	1398	3	3
Pair 7a-7b	1511	1623	4	4
Pair 8a-8b	1234	1329	2	2
Pair 9a-9b	1555	1567	1	1
Pair 10a-10b	2385	2447	2	2

4 Results of the Experimental Tests of User Identification Based on Fingerprint Analysis

In order to acknowledge the effectiveness of user identification based on fingerprint analysis with the use of the [Q] measure indication, the authors of the article conducted a series of experimental tests. The method proposed by the authors of the article was compared with a popular and very effective method of fingerprint identification, the so called "*minutiae examination method*" [1, 10, 11]. The tests did not use fingerprints available in public data bases such as e.g. Fingerprint Verification Competition 2004—FVC 2004 [12]. The authors of the article created their own data base containing digital images of fingerprints of 127 users. The users' fingerprints stored in the data base constitute reference models for comparison in the identification process. The models stored in the base were compared with fingerprints taken on a current basis, whereas each taking may differ with translation and rotation with regard to the model stored in the base which constitutes a reference. All fingerprints were taken with the multi-function device HP Deskjet Advantage with the resolution of 200 dpi. The selection of this device was caused by the fact that it is very popular, cheap and the authors had such a device at their disposal. During the experimental tests, the authors of the article assumed that, in order to consider user identification proper, the digital fingerprint image taken on a current basis compared with the model stored in the base must have indications of the [Q] measure above 0.801 in 6 out of 9 segments. Otherwise, it is necessary to repeat the test and, if there is no positive result, it must be considered that the user has not been identified. The results of the experimental tests were presented in Table 8.

Analysing the example experimental results presented in Table 8, it is possible to conclude that the proposed method of user identification based on the [Q] measure indication analysis combined with the proposed synchronisation method based on the [Q] measure indication analysis is effective. The method proposed by the authors of the article is of comparable effectiveness as the method based on minutiae examination, but it is significantly faster.

Table 9 presents the results of average effectiveness and time of finding for both methods carried out for 127 users.

5 Conclusion

It can be concluded that the proposed method of user identification based on the [Q] measure indication analysis combined with the proposed synchronising algorithm both in the X and Y plane as well as taking into consideration the rotation is effective. The method may successfully be used in user identification systems. One could argue why the authors of the article used such a measure for objective assessment instead of a different one like e.g. the popular Mean Square Error MSE [9]. The selection was caused by the fact that the [Q] measure is relatively new and,

Table 8 Results of experimental tests concerning fingerprint recognition

Fingerprints	Required synchronisation	Number of segments where indication $Q > 0.901$	Assessment with the use of a method based on minutiae examination	Assessment with the use of a criterion	Repeated assessment with the use of a method based on minutiae examination	Number of segments in repeated measurement where $Q > 0.901$	Repeated assessment with the use of a criterion
Pair 1	Yes	6	The same	The same	X	X	
Pair 2	Yes	6	Different	The same	The same	X	
Pair 3	Yes	8	The same	The same	X	X	
Pair 4	Yes	9	The same	The same	X	X	
Pair 5	Yes	8	The same	The same	X	X	
Pair 6	Yes	5	The same	Different	X	7	The same
Pair 7	Yes	4	The same	Different	X	6	The same
Pair 8	Yes	6	The same	The same	X	X	
Pair 9	Yes	3	Different	Different	Different	3	Different
Pair 10	Yes	7	The same	The same	X	X	

Table 9 Comparing of the effectiveness of the methods for 127 users

	Average effectiveness (%)	Average effectiveness for for 1st trial 2nd trial (%)	Average identification time (s)
Method Based on [Q] indication analysis	67	100	7
Method of minutiae examination	68	100	13

moreover, the measure and its modifications are being constantly developed by introducing new and more effective modifications.

Obviously, the proposed synchronisation algorithm may also be used in other areas like e.g. in the process of creating panoramic photographs [13].

References

1. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of fingerprint recognition. Originally published in the series: Springer Professional Computing 2nd ed. (2009)
2. Gutkowska, D., Stolc, L.: Techniques of Identifying Individuals with the Use of Individual Biometric Features. Scientific Journals of the Faculty of Electronics and Automation Gdansk University of Technology No. 20 (in polish) (2004)
3. Swiatek, J.: Two-stage Identification and its technical and biomedical applications. Politechnic Warsaw Editors, Warsaw, Poland (in polish) (1997)
4. Maltoni, D., Maio, D.: Direct gray-scale minutiae detection in fingerprints. IEEE Trans. PAMI **19**, 27–40 (1997)
5. Sandstrom, M.: Liveness detection in fingerprint recognition systems. Master thesis (2004). http://www.ep.liu.se/exjobb/isy/2004/3557/exjobb.pdf
6. Tadeusiewicz, R., Korohoda, P.: Computer analysis and image processing. Fundacja Postępu Telekomunikacji (in polish) (1997)
7. Kornatowski, E., Kowalski, J., Mikolajczak, G., Peksinski J.: Linear and non-linear filtration of discrete images. In: Hogben (ed.) Szczecin, Poland (in polish) (2006)
8. Peksinski, J., Mikolajczak, G.: Using a neural network to generate a FIR filter to improves digital images using a discrete convolution operation. Intelligent information and database systems (ACIIDS 2012), LNAI 7197, pp. 294–300. Springer (2012)
9. Peksinski, J., Mikolajczak, G.: The synchronization of the images based on normalized mean square error algorithm. In: Nguyen NT (ed.) Advances in Multimedia and Network Information System Technologies Book Series: Advances in Intelligent and Soft Computing, vol. 80, pp. 15–25 (2010)
10. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Processing Lett. **9**(3), 81–84 (2002)
11. Maio, D., Maltoni, D.: Direct gray-scale minutiae detection in fingerprints (1997). http://bias.csr.unibo.it/fvc2004/
12. Wang, Z., Bovik, A.C.: Mean squared error: love it or leave it? A new look at signal fidelity measures. IEEE Signal Process. Mag. **26**(1), 98–117 (2009)
13. Kowalski, J., Pęksinski, J., Mikolajczak, G.: Using the Q measure to create panoramic photographs. In: 38th International Conference on Telecommunications and Signal Processing, pp. 560-563, IEEE Press, New York (2015)

A Compound Moving Average Bidirectional Texture Function Model

Michal Haindl and Michal Havlíček

Abstract This paper describes a simple novel compound random field model capable of realistic modelling the most advanced recent representation of visual properties of surface materials—the bidirectional texture function. The presented compound random field model combines a non-parametric control random field with local multispectral models for single regions and thus allows to avoid demanding iterative methods for both parameters estimation and the compound random field synthesis. The local texture regions (not necessarily continuous) are represented by an analytical bidirectional texture function model which consists of single scale factors modeled by the three-dimensional moving average random field model which can be analytically estimated as well as synthesized.

Keywords Bidirectional texture function · Texture synthesis · Compound random field model

1 Introduction

Convincing and physically correct virtual models require not only precise 3D shapes in accord with the captured scene, but also object surfaces covered with genuine nature-like surface material textures with physically correct reflectance to ensure realism in virtual scenes. The primary purpose of any synthetic texture approach is to reproduce and enlarge a given measured texture image so that ideally both natural and synthetic texture will be visually indiscernible. However, the appearance of real materials dramatically changes with illumination and viewing variations. Thus, the only reliable representation of material visual properties requires capturing of its reflectance in as wide range of light and camera position combinations as possible.

M. Haindl (✉) · M. Havlíček
Institute of Information Theory and Automation of the CAS, Prague, Czech Republic
e-mail: haindl@utia.cz

M. Havlíček
e-mail: havlimi2@utia.cz

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_8

This is a principle of the recent most advanced texture representation, the seven dimensional Bidirectional Texture Function (BTF) [13]. Compound random field models consist of several sub-models each having different characteristics along with an underlying structure model which controls transitions between these sub models [19]. Compound Markov random field models (CMRF) were successfully applied to image restoration [3, 5, 19, 21], segmentation [24], or modeling [9, 11, 16, 17]. However, these models always require demanding numerical solutions with all their well known drawbacks. The exceptional CMRF [9] model allows analytical synthesis at the cost of a slightly compromised compression rate.

We propose a compound moving average bidirectional texture function model BTF-CMA model which combines a non-parametric and parametric analytically solvable moving average (MA) random fields (RF) and thus we can avoid using some of time consuming iterative Markov Chain Monte Carlo (MCMC) method for both BTF-CMA model parameters estimation as well as BTF-CMA synthesis. Similarly to the previously mentioned CMRF methods, our presented model avoids range map estimation which is required for most RF based BTF models [8, 12, 14, 15, 18]. Beside texture synthesis, texture editing is another useful application which has large potential for significant speed-up and cost reduction in industrial virtual prototyping [16]. Although some recent attempts have been made to automate this process, automatic integration of user preferences still remains an open problem in the context of texture editing [16]. Proposed method present partial solution of this problem by combining estimated local models from several different source textures or simply editing estimated local models of the original texture.

2 Compound Random Field Texture Model

Let us denote a multiindex $r = (r_1, r_2)$, $r \in I$, where I is a discrete 2-dimensional rectangular lattice and r_1 is the row and r_2 the column index, respectively. $X_r \in \{1, 2, \dots, K\}$ is a random variable with natural number value (a positive integer), Y_r is multispectral pixel at location r and $Y_{r,j} \in \mathscr{R}$ is its j-th spectral plane component. Both random fields (X, Y) are indexed on the same lattice I. Let us assume that each multispectral or BTF observed texture $\tilde{Y}$ (composed of d spectral planes) can be modelled by a compound random field model, where the principal random field X controls switching to a regional local model $Y = \bigcup_{i=1}^{K} {}^{i}Y$. Single K regional submodels ${}^{i}Y$ are defined on their corresponding lattice subsets ${}^{i}I$, ${}^{i}I \cap {}^{j}I = \emptyset \quad \forall i \neq j$ and they are of the same RF type. They differ only in their contextual support sets ${}^{i}I_r$ and corresponding parameters sets ${}^{i}\theta$. The CRF model has posterior probability

$$P(X, Y \mid \tilde{Y}) = P(Y \mid X, \tilde{Y}) P(X \mid \tilde{Y})$$

and the corresponding optimal MAP solution is:

$$(\hat{X}, \hat{Y}) = \arg \max_{X \in \Omega_X, Y \in \Omega_Y} P(Y \mid X, \tilde{Y})\, P(X \mid \tilde{Y}) \ ,$$

where Ω_X, Ω_Y are corresponding configuration spaces for random fields (X, Y).

2.1 Region Switching Model

The principal RF $(P(X \mid \tilde{Y}))$ can be, for example, represented by a flexible $K-$state Potts random field [17, 22, 23]. Instead of the Potts RF or some alternative general parametric MRF, which require a Markov chain Monte Carlo (MCMC) solution, we suggest to use simple non-parametric approximation based on our roller method [6, 7].

The control random field $\breve{X}$ is estimated using simple K-means clustering of $\tilde{Y}$ in the RGB colour space into predefined number of K classes, where cluster indices are $\breve{X}_r \quad \forall r \in I$ estimates. The number of classes K can be estimated using the Kullback-Leibler divergence and considering sufficient amount of data necessary to reliably estimate all local Markovian models.

The roller method is subsequently used for optimal $\breve{X}$ compression and extremely fast enlargement to any required field size. The roller method [6, 7] is based on the overlapping tiling and subsequent minimum error boundary cut. One or several optimal double toroidal data patches are seamlessly repeated during the synthesis step. This fully automatic method starts with the minimal tile size detection which is limited by the size of control field, the number of toroidal tiles we are looking for and the sample spatial frequency content. The roller method advantageously maintains the original overall ratio single regions areas, e.g., the average standard deviation for this percentage ratio after four times enlarged texture map was observed to be less than 3 %.

2.2 Spatial Factorization

The spatial factorisation is technique that enables separate modelling of individual band limited frequency components of input image data and thus to use random field models with small compact contextual support. This factorization step is the prerequisite for satisfactory visual quality result of the presented model. Each grid resolution represents a single spatial frequency band of the texture which corresponds to one layer of Gaussian pyramid [13]. The input data are decomposed into a multi-resolution grid and all resolution data factors represents the Gaussian-Laplacian pyramid of level k which is a sequence of k images in which each one is a low-pass down-sampled version of its predecessor.

2.3 Local Moving Average Models

Single multispectral texture factors are modelled using the extended version ($3D\,MA$) of the moving average model [20]. A stochastic multispectral texture can be considered to be a sample from 3D random field defined on an infinite 2D lattice. A spatial input factor Y is represented by the $3D\,MA$ random field model. Y_r is the intensity value of a multispectral pixel $r \in I$ in the image space. The model assumes that each factor is the output of an underlying system which completely characterizes it in response to a 3D uncorrelated random input. This system can be represented by the impulse response of a linear 3D filter. The intensity values of the most significant pixels together with their neighbours are collected and averaged, and the resultant 3D kernel is used as an estimate of the impulse response of the underlying system. A synthetic mono-spectral factor can be generated by convolving an uncorrelated 3D random field with this estimate. Suppose a stochastic multi-spectral texture denoted by Y is the response of an underlying linear system which completely characterizes the texture in response to a 3D uncorrelated random input E_r, then Y_r is determined by the following difference equation:

$$Y_r = \sum_{s \in I_r} B_s E_{r-s} \tag{1}$$

where B_s are constant matrix coefficients and $I_r \subset I$. Hence Y_r can be represented $Y_r = h(r) * E_r$ where the convolution filter $h(r)$ contains all parameters B_s. In this equation, the underlying system behaves as a 3D filter, where we restrict the system impulse response to have significant values only within a finite region. The geometry of I_r determines the causality or non-causality of the model. The selection of an appropriate model support region is important to obtain good results: small ones cannot capture all details of the texture and contrariwise, inclusion of the unnecessary neighbours adds to the computational burden and can potentially degrade the performance of the model as an additional source of noise.

The parameter estimation can be based on the modified Random Decrement technique (RDT) [1, 2]. RDT assumes that the input is an uncorrelated random field. If every pixel component is higher than its corresponding threshold vector component and simultaneously at least one of its four neighbours is less than this threshold the pixel is saved in the data accumulator. The procedure begins by selecting thresholds usually chosen as some percentage of the standard deviation of the intensities of each spectral plane separately. Additionally to that, a 3D MA model requires also to estimate the noise spectral correlation, i.e.,

$$\begin{aligned} &E\{E_r E_s\} = 0 && \forall r_1 \neq s_1 \vee r_2 \neq s_2 \ , \\ &E\{E_{r_1,r_2,r_3} E_{r_1,r_2,\bar{r}_3}\} \neq 0 && \forall r_3 \neq \bar{r}_3 \ . \end{aligned}$$

The synthetic factor can be generated simply by convolving an uncorrelated 3D RF E with the estimate of B according to (1). All generated factors form new

Gaussian pyramid. Fine resolution synthetic smooth texture is obtained by the collapse of the pyramid i.e. an inverse procedure of that one creating the pyramid.

The resulting synthesized texture is obtained by mapping individual synthesized local sub textures to the enlarged control field realization. Additional pixel swapping and filtering along the individual region border increases the visual quality of the result as the overall intensity of the borders may be distracting.

3 Results

Automatic texture quality evaluation is important but still unsolved difficult problem and qualitative evaluation is for now possible only using impractical and expensive visual psycho-physics. We have recently tested [10] on our texture fidelity benchmark

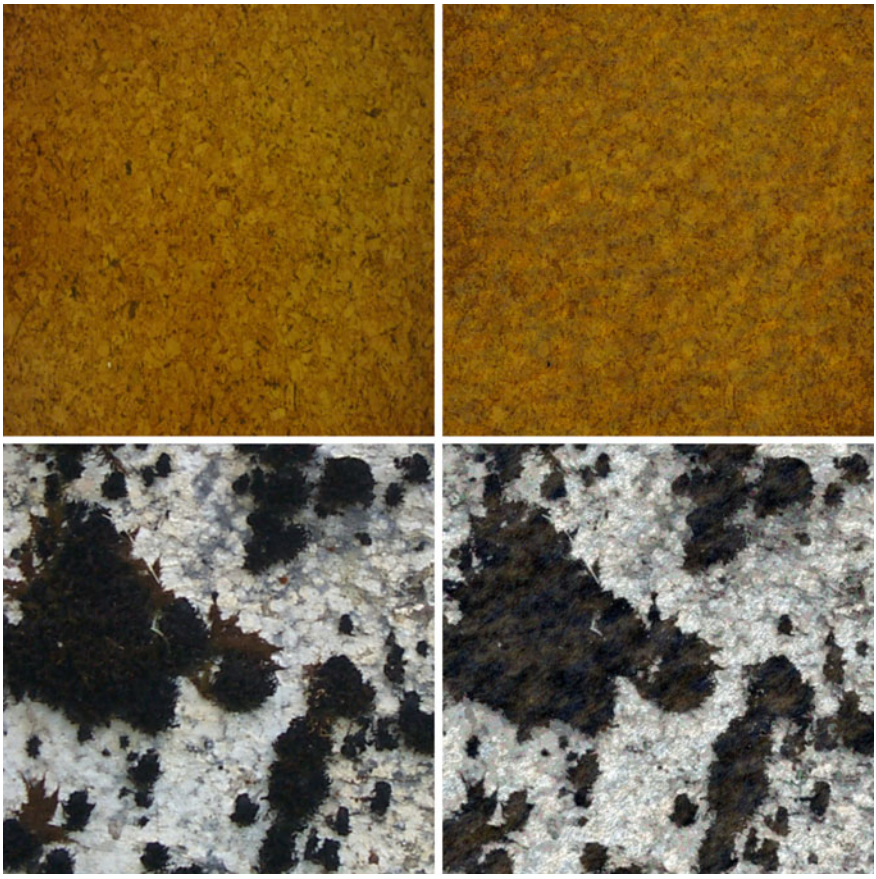


Fig. 1 Examples of the cork texture (*upper row*) and lichen (*bottom row*) and their CMRF^{3MA} synthesis (*right column*)

(http://tfa.utia.cas.cz) several published state-of-the-art image quality measures and also one dedicated texture measure (STSIM) in several variants. We have tested the presented novel $BTF - CMRF^{3MA}$ model on natural colour textures from our extensive texture database (http://mosaic.utia.cas.cz), which currently contains over 1000 colour or BTF textures. Tested textures were either natural, such as two textures on Figs. 1, 2, 3, 5 or man-made Fig. 4 (terracotta). Tested BTF material samples from our database [4] are measured in 81 illumination and viewing angles, respectively. A material sample measurements (Fig. 4) from this database have resolution of 1800×1800 and size 1.2 GB. Figure 4 shows a cutout example from such measurements of a terracotta material and its synthesis for two different illumination and view angle combinations. All presented examples use five level control field ($K = 5$), the hierarchical contextual neighbourhood of the third order, and the three-layer Gaussian-Laplacian pyramid.

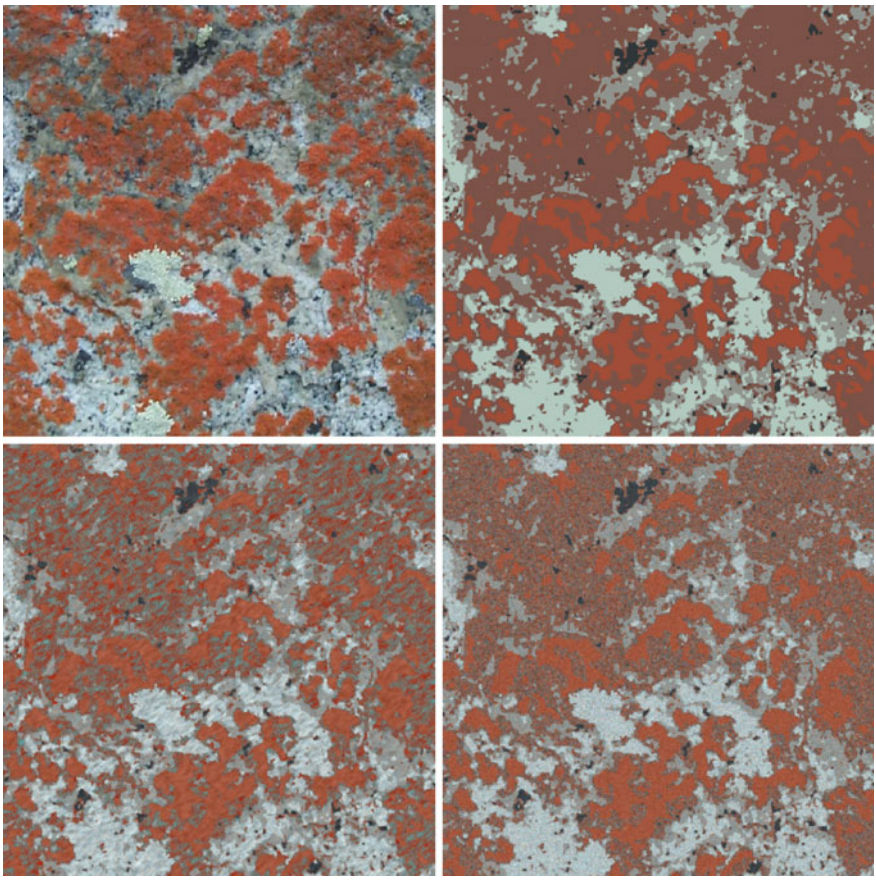


Fig. 2 An example of the lichen texture (*upper left*), its control field (*upper right*), the CMRF^{3MA} synthesis (*bottom left*), and a comparative synthesis using a 3D Gaussian generator

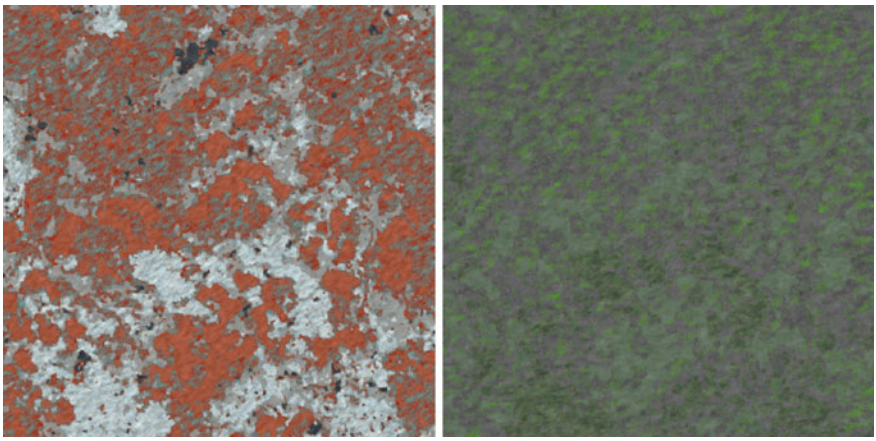


Fig. 3 Synthetic (CMRF^{3MA}) lichen texture and its edited version (*right*)

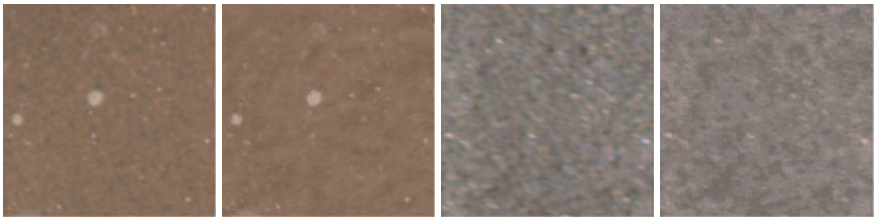


Fig. 4 An example of the measured BTF terracotta texture and its synthetic (even images) results, where ϕ, θ are azimuthal and elevation illumination/viewing angles, respectively

Figure 2 advantageously compares the presented $BTF - CMRF^{3MA}$ model (Fig. 2-bottom left) with local fields modeled by simple multidimensional Gaussian generator (Fig. 2-bottom right). The Gaussian generator produces too noisy and spatially uncorrelated synthetic texture (e.g. top right corner). The model can be easily used to create an artificial texture by editing single local sub-textures (Fig. 3), which can be either learned from separate sources or their parameters can be manually modified. Figure 5 illustrates a fourfold enlarged stone texture.

Resulting synthetic more complex textures (such as lichen on Figs. 1-bottom, 2) have generally better visual quality (there is no any usable analytical quality measure) than textures synthesised using our previously published [8, 12, 15] simpler MRF models. Synthetic multispectral textures are mostly surprisingly good for such a fully automatic fast algorithm. Obviously there is no universally optimal texture modelling algorithm and also the presented method will produce visible repetitions for textures with distinctive low frequencies available in small patch measurements (relative to these frequencies). BTF-CMRF is capable to reach huge BTF compression ration $\sim 1{:}1 \times 10^5$ relative to the original BTF measurements but $\approx 5\times$ lower than [12].

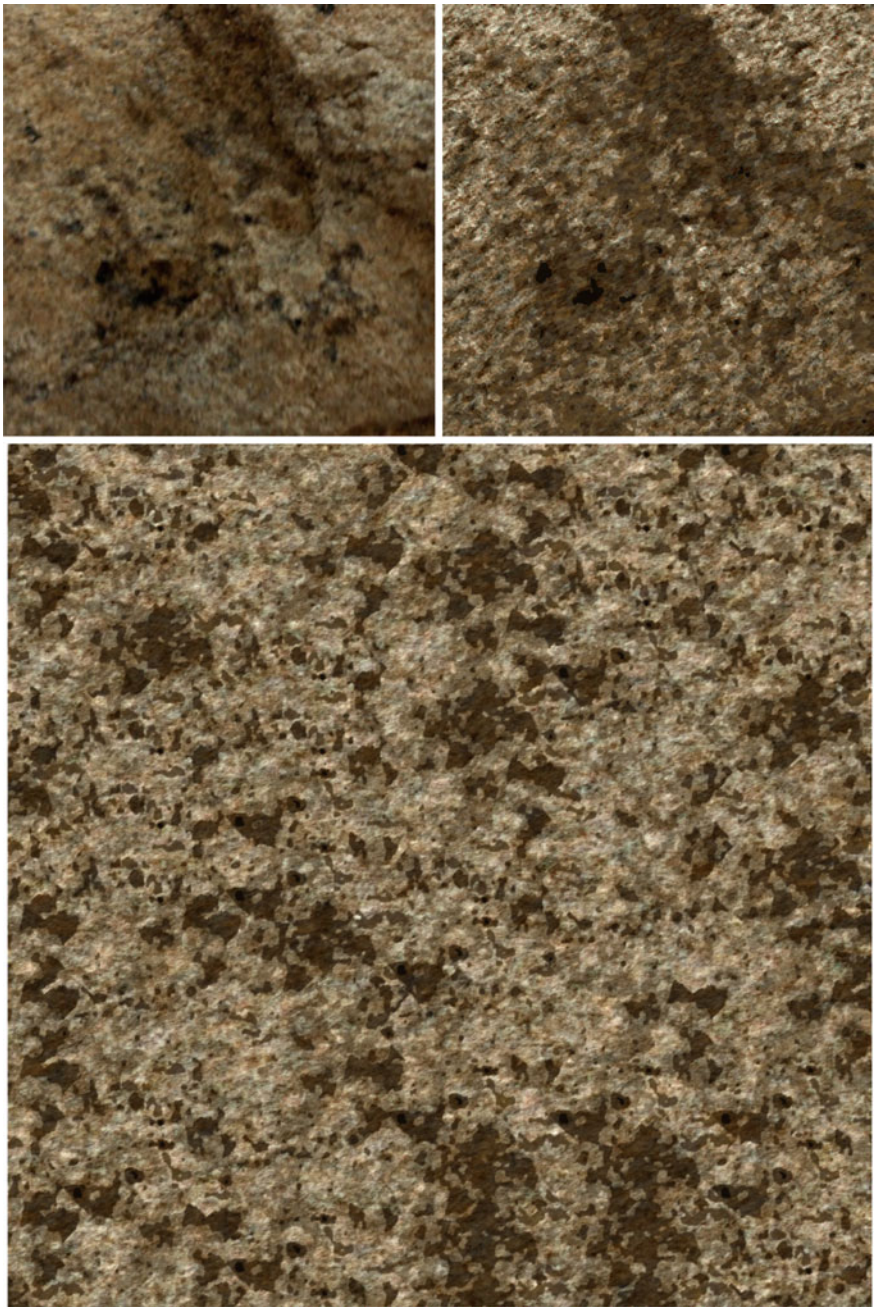


Fig. 5 An example of the stone texture (*upper left*), its original size (*upper right*) and fourfold enlarged CMRF^{3MA} synthesis

4 Conclusions

The presented CMRF (BTF-CMRF) method shows good visual performance on selected real-world materials. The appearance of such materials should consist of several types of relatively small regions with fine-granular inner structure such as sand, grit, cork, lichen, or plaster. The model offers large data compression ratio (only tens of parameters per BTF and few small control field tiles) easy simulation and exceptionally fast seamless synthesis of any required texture size. The method can be easily generalised for colour or BTF texture editing by estimating some local models on one or several target textures. Both analysis as well as synthesis of the model are exceptionally fast. The model does not compromise spectral correlation thus it can reliably model motley textures. A drawback of the method is that it does not allow a BTF data space restoration or modelling of unseen (unmeasured) BTF space data unlike some fully parametric probabilistic BTF models, and it requires a pyramidal spatial factorization.

Acknowledgments This research was supported by the Czech Science Foundation project GAČR 14-10911S.

References

1. Asmussen, J.C.: Modal analysis based on the random decrement technique: application to civil engineering structures. Ph.D. thesis, University of Aalborg (1997)
2. Cole Jr, H.A.: On-line failure detection and damping measurement of aerospace structures by random decrement signatures. Technical Report TMX-62.041, NASA (1973)
3. Figueiredo, M., Leitao, J.: Unsupervised image restoration and edge location using compound Gauss–Markov random fields and the mdl principle. IEEE Trans. Image Process. **6**(8), 1089–1102 (1997)
4. Filip, J., Haindl, M.: Bidirectional texture function modeling: a state of the art survey. IEEE Trans. Pattern Anal. Mach. Intell. **31**(11), 1921–1940 (2009)
5. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions and bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**(11), 721–741 (1984)
6. Haindl, M., Hatka, M.: BTF Roller. In: Chantler, M., Drbohlav, O. (eds.) Texture 2005. Proceedings of the 4th International Workshop on Texture Analysis. pp. 89–94. IEEE, Los Alamitos (2005)
7. Haindl, M., Hatka, M.: A roller—fast sampling-based texture synthesis algorithm. In: Skala, V. (ed.) Proceedings of the 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, pp. 93–96. UNION Agency—Science Press, Plzen (2005)
8. Haindl, M., Havlíček, V.: A multiscale colour texture model. In: Kasturi, R., Laurendeau, D., Suen, C. (eds.) Proceedings of the 16th International Conference on Pattern Recognition. pp. 255–258. IEEE Computer Society, Los Alamitos (2002). http://dx.doi.org/10.1109/ICPR.2002.1044676
9. Haindl, M., Havlíček, V.: A compound MRF texture model. In: Proceedings of the 20th International Conference on Pattern Recognition, ICPR 2010. pp. 1792–1795. IEEE Computer Society CPS, Los Alamitos (2010). http://doi.ieeecomputersociety.org/10.1109/ICPR.2010.442

10. Haindl, M., Kudělka, M.: Texture fidelity benchmark. In: 2014 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM), pp. 1–5. IEEE Computer Society CPS, Los Alamitos (2014)
11. Haindl, M., Remeš, V., Havlíček, V.: Potts compound markovian texture model. In: Proceedings of the 21st International Conference on Pattern Recognition. ICPR 2012, pp. 29–32. IEEE Computer Society CPS, Los Alamitos (2012)
12. Haindl, M., Filip, J.: Extreme compression and modeling of bidirectional texture function. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1859–1865 (2007). http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.1139
13. Haindl, M., Filip, J.: Visual texture. Advances in Computer Vision and Pattern Recognition. Springer, London (2013)
14. Haindl, M., Havlíček, M.: Bidirectional texture function simultaneous autoregressive model. In: Salerno, E., Etin, A., Salvetti, O. (eds.) Computational Intelligence for Multimedia Understanding, Lecture Notes in Computer Science, vol. 7252, pp. 149–159. Springer, Berlin (2012). doi:10.1007/978-3-642-32436-9_13. http://www.springerlink.com/content/hj32551334g61647/
15. Haindl, M., Havlíček, V.: A multiresolution causal colour texture model. Lect. Notes Comput. Sci. **1876**, 114–122 (2000)
16. Haindl, M., Havlíček, V.: A plausible texture enlargement and editing compound markovian model. In: Salerno, E., Cetin, A., Salvetti, O. (eds.) Computational Intelligence for Multimedia Understanding, Lecture Notes in Computer Science, vol. 7252, pp. 138–148. Springer, Berlin (2012). doi:10.1007/978-3-642-32436-9_12. http://www.springerlink.com/content/047124j43073m202/
17. Haindl, M., Remeš, V., Havlíček, V.: Btf potts compound texture model, vol. 9398, pp. 939807-1–939807-11. SPIE, Bellingham, WA 98227-0010, USA (2015). http://dx.doi.org/10.1117/12.2077481
18. Havlíček, M., Haindl, M.: A moving average bidirectional texture function model. In: Wilson, R., Hancock, E., Bors, A., Smith, W. (eds.) Computer Analysis of Images and Patterns. Lecture Notes in Computer Science, vol. 8048, pp. 338–345. Springer (2013)
19. Jeng, F.C., Woods, J.W.: Compound Gauss-Markov random fields for image estimation. IEEE Trans. Signal Process. **39**(3), 683–697 (1991)
20. Li, X., Cadzow, J., Wilkes, D., Peters, R., II Bodruzzaman, M.: An efficient two dimensional moving average model for texture analysis and synthesis. In: Proceedings IEEE Southeastcon '92, vol. 1, pp. 392–395. IEEE (1992)
21. Molina, R., Mateos, J., Katsaggelos, A., Vega, M.: Bayesian multichannel image restoration using compound Gauss-Markov random fields. IEEE Trans. Image Proc. **12**(12), 1642–1654 (2003)
22. Potts, R., Domb, C.: Some generalized order-disorder transformations. Proc. Cambr. Philos. Soc. **48**, 106–109 (1952)
23. Wu, F.: The Potts model. Rev. Modern Phys. **54**(1), 235–268 (1982)
24. Wu, J., Chung, A.C.S.: A segmentation model using compound markov random fields based on a boundary model. IEEE Trans. Image Process. **16**(1), 241–252 (2007)

Part II
Voice Interactions in Multimedia Systems

Multiple Information Communication in Voice-Based Interaction

Muhammad Abu ul Fazal and M. Shuaib Karim

Abstract Ubiquitous Computing has enabled users to perform their computer activities anytime, anyplace, anywhere while performing other routine activities. Voice-based interaction often plays a significant role to make this possible. Presently, in voice-based interaction system communicates information to the user sequentially whereas users are capable of noticing, listening and comprehending multiple voices simultaneously. Therefore, providing information sequentially to the users may not be an ideal approach. There is a need to develop a design strategy in which information could be communicated to the users through multiple channels. In this paper, a design possibility has been investigated that how information could be communicated simultaneously in voice-based interaction so that users could fulfil their growing information needs and ultimately complete multiple tasks at hand efficiently.

Keywords Voice-based interaction · Multiple information broadcast · Multiple voices · Information design

1 Introduction

In this information age which is highly influenced by technology, people have many computing devices and associated interaction modes to fulfill information needs and perform desired tasks conveniently from anywhere [1]. For example, mobile telephony has become an essential tool [2] that humans carry with them almost all the time. It is playing a significant role to access information on the go by either interacting visually or by using voice-based interaction. A voice-based interaction is a mode where users are provided with the facility to interact with the system using 'voice'.

M.A. ul Fazal (✉) · M.S. Karim (✉)
Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Pakistan
e-mail: fazalsidhu@yahoo.com

M.S. Karim
e-mail: skarim@qau.edu.pk

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_9

The motivation of using voice to interact with the system is an old concept which can be associated with Ali Baba's 'Open Sesame' and earlier science fiction movies. The voice-based interaction method enables the user to interact with the system in immersive environment [3]. *Voice User Interfaces (VUI)s are user interfaces using speech input through a speech recognizer and speech output through speech synthesis or pre-recorded audio* [4].

The voice-based interaction enables users to conveniently interact with the system in the hand busy or the eye busy environment. This mode is also an alternative for the visually impaired users to interact with systems. According to world health organization [5], it is estimated that there are 285 million people who have visual impairments.

Humans are capable of listening and comprehending multiple information simultaneously through their auditory perception, but presently, voice-based interaction design is providing sequential interaction approach which is somehow under-utilizing the natural human perception capabilities [6]. Since the voice-based interaction is sequential therefore system provides only a limited amount of information each time, which makes it hard for the users to get an overview of the information, particularly in the case of assistive technology used by the visually impaired users [7].

One of the main goals in information design is rapid dissemination with clarity. Since users have growing information needs [8], therefore, it must be efficiently designed, produced and distributed, so that users could quickly interpret and understand it using their auditory capabilities. If we critically look contemporary implementations, then there arises a question whether present sequential information designs are utilizing the human auditory capabilities in voice-based interaction effectively & optimally or not?

Rest of the paper is organized as follows. Next section describes Literature Review. The limited exploitation of human auditory perception is discussed under the section Auditory Perception's Exploitation Gap. The concept of communicating multiple information using multiple voice streams is discussed under Motivating Scenario section. Then, based on the motivating scenario, an experiment is described in detail under Experiment Section. Conclusion and future work is discussed under Conclusion and Future Work section.

2 Literature Review

Voice-based interactions often used in todays' computing era that is ubiquitous in nature. Over the Web, efforts have also been made to realize voice-based user agents such as voice-based Web browsers under the Spoken Web Initiative [9]. That would benefit people who are unable to conveniently use the internet due to various reasons including low literacy, poverty, and disability.

There are many other uses of voice in system interaction like in e-learning system, the aural access is being provided as a complimentary method to the visual-only content [10]. Numerous interactive voice response applications are developed to provide

important information to the targeted users, particularly the illiterate users. Interactive voice application 'Avaaj Otalo' [11] provides essential information to the low literate rural formers. Using this application, farmers can ask questions, and browse stored responses on a range of agricultural topics.

From the user side, Lewis suggested that user system interaction performance is affected by the users' characteristics like physical, mental, and sensory abilities [12]. For voice, the main sensory capability is auditory acuity. The American Speech-Language-Hearing Association has identified central auditory process as the auditory system mechanisms and processes responsible for the following behaviors [13]:

- Sound localization and lateralization, i.e. users are capable of knowing the space where sound has occurred
- Auditory discrimination, i.e. user has the ability to distinct one sound from another
- Auditory pattern recognition, i.e. user is capable of judging differences and similarities in patterns of sounds
- Temporal aspects, i.e. user has abilities to sequence sounds, integrate a sequence of sounds into meaningful combinations, and perceive sounds as separate when they quickly follow one another
- Auditory performance decrements, i.e. user is capable of perceiving speech or other sounds in the presence of another signal
- Auditory performance with degraded acoustic signals, i.e. user has the ability to perceive a signal in which some of the information is missing

Humans are able to listen to the sound whose frequency varies between 16 Hz to 20 KHz. In order to perceive the two frequencies separately the width of the filters, also called 'critical band', determines the minimum frequency spacing. It would be difficult to separate two sounds if it falls within the same critical band. Besides frequency, other important perceptual dimensions are pitch, loudness, timbre, temporal structure and spatial location.

Humans are capable of focusing their attention to an interested voice stream if they perceive multiple information simultaneously as reflected in experiment discussed in this paper. For attention user adopts two kinds of approaches, one is overt attention and second is covert attention. In covert attention the region of interest is in the periphery. So, if a user is listening multiple voices, he may be interested in focusing the voice provided to him in the periphery. The regions of interest could be four to five [14]. For selection and attention in competing sounds, it is an important consideration for the listener that how auditory system organizes information into perceptual 'streams' or 'objects' when multiple signals are sent to the user. In order to meet this challenge, auditory system groups acoustic elements into streams, where the elements in a stream are likely to come from the same object (Bregman 1990).

A few research studies exist on communicating information using voice simultaneously. The experiments have been conducted particularly in the case of visually impaired persons. According to Guerreiro, multiple simultaneous sound sources can help blind users to find information of interest quicker by scanning websites with several information items [15]. Another interesting work where Hussain introduced

hybrid feedback mechanism i.e. speech based and non-speech based (spearcon) feedback to the visually impaired persons while they travel towards their destination [16]. The feedback mode alters between above two modes on the basis of the frequency of using the same route by the user and representativeness of the same feedback provided to the user. The experiment conducted by the researcher reflects that hybrid feedback is more effective than the speech only feedback and non-speech only feedback. In another study for blinds to understand in a better way the relevant source's content, Guerreiro and Goncalves, established that use of two to three simultaneous voice sources provide better results [17]. The increasing number of simultaneous voices decreases the source identification and intelligibility of speech. Secondly, the author found that the location of sound source is the best mechanism to identify content.

Above mentioned behavioral characteristics and research work suggest that human auditory perception has remarkable capabilities which are somehow not fully exploited in the contemporary implementations of the voice-based human-computer interaction, particularly for sighted users.

3 Suggested Improvements

Contrary to the voice-based interface, the visual interface provides multiple information to the user in many ways such as using overlays [18]. Figure 1 is a Facebook wall of a user where multiple information is being communicated simultaneously. One overlay is providing the facility to view the messages being received in the

Fig. 1 An example of overlay in Graphical User Interface

conversation. Another overlay at the top is showing notifications. The right side pane is showing the activities of fellows. The left side pane is displaying his favorites and other useful stuff. And as soon as the mouse is rolled over to the text Farrukh Tariq Sidhu the preview of Farrukh's wall gets displayed in another layer. If the user is interested in the additional information provided through overlay the user may go with it otherwise ignore the overlay and would stay on the main screen.

The same design technique may be adopted in voice response system to communicate multiple information simultaneously because auditory system is capable of performing filtration of received sounds and allows the user to ignore the irrelevant noise and concentrate on important information [19].

In next section, we have discussed a scenario where multiple voice streams can help users to fulfill their information needs.

4 Motivating Scenario: Listening Multiple Talk Shows

Daily, in prime time i.e. 8:00 pm to 9:00 pm various news channels air talk shows focusing different topics with different participants and hosts. People working in offices in evening or night shifts usually watch these programs live using video streams provided by news channels, if they are free to do so at the desk. If users are busy in official work or their computer screen is occupied for another task they may prefer to listen to live audio stream from relevant channels website.

Users may be interested in listening to more than one talk shows at the same time. For an example, a person is interested in listening to the talk show 'Capital Talk' at 'Geo News' and also interested in listening to 'Off the Record' played on another channel 'ARY News'. The first talk show Capital Talk is discussing the current situation arisen due to the heavy floods whereas the second program is discussing the political scenario in Pakistan. The user is mainly interested in listening to the program discussing the political situation but also wants to know the key facts or get an overview about the flood situation being discussed in Capital Talk show.

In this perspective, user's multiple information needs may be fulfilled using multiple information communication simultaneously. In this case, information seeking could be possible in a way that a user opens two web browsers and play both the audio streams simultaneously and listens both the program in parallel. This could be challenging and complex task for the user. The listening complexity may be reduced by keeping one streams volume low but audible and keep the main programs voice normal so that user could keep the focus on primary program. The high volume is expected to help him to keep the focus on the main program while the secondary low volume would continuously give him the feedback or glimpse that what is going on in the other program. Using this approach user might not miss the content of the program in which user is mainly interested and also get an overview of the secondary program.

This approach of playing multiple audio streams in parallel may be extended to more than two audio streams where information like a commentary on cricket match could also have listened.

In order to meet this challenge, we have framed following three research questions which we are trying to answer by conducting a series of experiments.

- How many voice streams can optimally be played to users for communicating information simultaneously?
- What could be the optimal auditory perceptual dimensions' settings of streams for better discrimination between voice streams?
- What scenarios/challenges users can face in multiple information communication?

5 Experiment

In this experiment, an audio bulletin was built wherein the voice-based information was designed in a way that two different voice streams (using female and male voices) were played simultaneously. The female voice stream was of BBC Urdu's renowned TV presenter 'Aaliya Nazki' and reporter 'Nasreen Askri' whereas male voice was of another BBC Urdu's TV presenter 'Shafi Taqqi Jami'.

5.1 Experiment Design and Settings

In order to build an audio bulletin, two different video bulletin of BBC Urdu's program 'Sairbeen' were selected. Sairbeen is one of the renowned news bulletins that includes worldwide reports, expert opinions, public opinions, features on interesting topics and current affairs. This program is very popular among the public. These video bulletins were converted into two audio files of wav format. Each audio file consisted of three different news stories. From the first audio file which was in Aalia Nazki's voice, a detailed news about an exhibition scheduled to be held in Mohatta palace was selected. And from the second audio file of Shafi Taqi Jami, the main headlines of all three news were selected. These three headlines were further broken into three audio files. Each audio file played a news headline.

In order to play these news streams a different information design strategy i.e. multiple information communication simultaneously was used. In this bulletin, the detailed exhibition news was set to play continuously throughout the bulletin in a female voice. This voice stream was termed as a primary voice in the experiment. Moreover, while keeping the primary voice in playing mode the other three news stored in three audio files were also played after periodic intervals of 10 s. This voice considered as a secondary voice. The primary voice was set to come from left earphone whereas the secondary voice was set to come from right earphone. This approach was adopted because it was expected that playing primary and secondary

voice in different earphones would bring ease for the user to discriminate both voice streams.

These two files with given information design were merged into one clip and played by writing a program in Visual Studio 2013 using C#. The total duration of this clip was 1 min and 28 s. This clip was played on Dell Vostro 5560 with Core i5 processor and 4GB RAM. In order to listen to the clip, an average quality KHM MX earphone was used to listen to the clip.

The experiment was conducted on people ranging from 20 to 55 years including both males and females. Total 10 users participated in this experiment out of which 6 were male and 4 were female. The experiment was conducted at random places without considering whether the environment/surrounding was fully quiet or not.

In order to judge the behavior of users in the experiment, a questionnaire was prepared. The interviewees were first briefed about the audio playing mechanism in this experiment. They were told about both the primary and the secondary voices. Before they started to listen to the audio clip they were given an overview of the questionnaire so that they could grab the information accordingly. The questionnaire aimed to establish whether a listener could notice, focus and comprehend multiple information simultaneously or not. It also helped to gauge the notice, selection and attention behavior of the user. In order to facilitate and reduce the memory load, participants were given maximum three choices to select one from.

5.2 Results

Most of the users were able to answer the questions correctly which were asked to find out, whether they could hear both the sounds simultaneously or not. And when they were asked about the perceptual and observational question all of the participants found voice streams audible, discriminable when played together.

Following is the response of users for each question asked in the questionnaire.

i. Could you hear the primary voice presenting documentary? From all participants, 80 % of the users told that the primary voice presenting documentary was clear. The remaining 20 % users who although said that they were able to listen to the primary voice but remarked that it was loud and shrilling so could further be improved.

ii. What was the topic of primary voice? All the participants rightly told the topic of primary voice i.e. Exhibition.

iii. Where was the exhibition scheduled to hold? 50 % of the users could not answer it correctly. Remaining those who answered correct, guessed it using their prior knowledge. The use of user's existing knowledge behavior would fully be investigated in upcoming series of experiments.

iv. What was the venue name? All the participants correctly answered the venue name of exhibition i.e. 'Mohatta Palace'.

v. Could you please tell us more about the exhibition documentary? In order to judge users' comprehension, they were asked to describe what they listened in

the exhibition documentary. All the users were able to describe the documentary and gave the overview in broken words. These words were kind of keywords in the documentary that users used.

It is observed that though users lost some amount of information while focusing on secondary voice but they still grasped the documentary very well and where there was an information gap they filled it with their existing knowledge.

vi. Could you notice the secondary voice? Yes, all users were able to notice the secondary voice in the presence of primary voice.

vii. Were you able to distinguish secondary voice in the presence of primary voice and vice versa? 70 % users stated that they found no difficulty to distinguish secondary sound from primary voice and vice versa. The 30 % users were of the view that it could further be improved.

It is learned that this easiness in discriminating both the voice streams was mainly possible, because, both sounds were coming in different ears separately and also the voice streams were uttered by different gender voices i.e. male and female. In order to make 'discrimination' more evident, other auditory dimensions could also be explored.

vii. Were you able to distinguish secondary voice in the presence of primary voice and vice versa? The 70 % users stated that they found no difficulty to distinguish secondary sound from primary voice and vice versa. Remaining 30 % users were of the view that it could further be improved because they missed some information while focusing a particular voice.

It is learned that this easiness in discriminating both the voice streams was mainly possible, because, both sounds were coming in different ears separately and also the voice streams were uttered by different gender voices i.e. male and female. In order the make 'discrimination' more evident, other auditory dimensions could be explored which we would do in future experiments.

viii. What was secondary voice indicating? All the users correctly answered that secondary voice was indicating news.

ix. How many times secondary voice played in different intervals? The 30 % users gave a wrong answer while 70 % users rightly told that it was played three times (Fig. 2).

The bar-chart indicates the number of correct/incorrect answers by the users for each question. The question v is discriptive, therefore, not reflected in bars whereas bars in question xiii indicate the selection of interested news by the users from three headlines played to them. In question i and vii the second blue bar indicate that how many users had asked to improve the quality.

x. In the first occurrence, what was the topic of secondary voice? Among all participants, only one user couldn't answer this question correctly.

xi. In the second occurrence, what was the topic of secondary voice? The 90 % of participants correctly told the topic of the second occurrence i.e. cyber attack.

xii. In the third occurrence, what was the topic of secondary voice? Same results were witnessed as seen in above two questions.

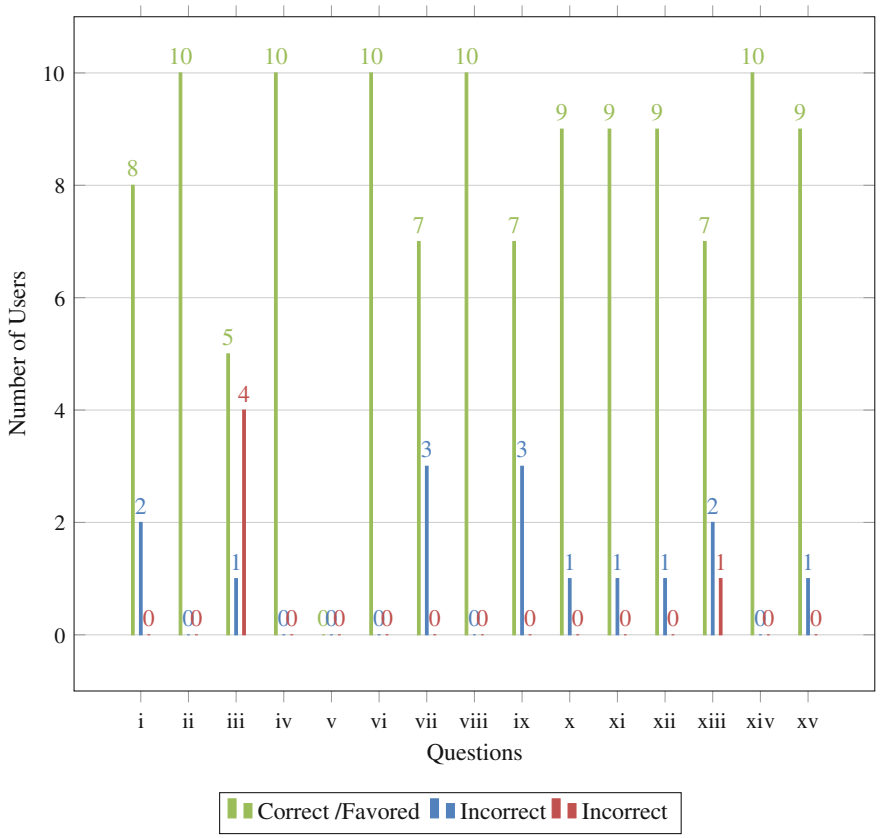


Fig. 2 Users' response performance in multiple voice-based information communication

xiii. Which was the most interesting news for you? The 70 % users opted 'Data theft in Cyber Attack' remaining 20 % of the users opted for 'Black Money in Budget' whereas only one female user showed interest in Exhibition Documentary.

xiv. Did you want to promptly listen to the detail news from any of the spoken news? As a follow-up to Question 13 when users asked to tell their intent that whether they wanted to promptly listen to the interesting news by skipping the present primary voice then 100 % of the users answered 'yes'.

This is an interesting finding which provides the opportunity of applying GUI based overlay, lightbox techniques in voice-based interaction which is discussed in the previous section using the Facebook wall of a user.

xv. Did you find multiple sounds helpful in reaching multiple information quickly and Would you prefer this approach over the sequential flow of information? The 90 % of the users found this quick design of delivering information helpful and said they would prefer this multiple information communication simultaneously over the sequential flow of information. From these 90 % users, a few had

reservations. They said, in this technique they are afraid that they might loose some important information which they would prefer to listen without any noise and disturbance. So, it could also be an interesting finding that in which contexts the multiple information communication design strategy could be applied and where it can't.

The 10 % of the users who didn't give preference said they are uni-task oriented so can't prefer this approach over the sequential flow of information.

6 Conclusion and Future Work

The results of this experiment are encouraging to further explore this design approach. The results validate that multiple information communication is possible using voice in Human-machine interaction. Users showed interest in multiple information communication. They were able to discriminate the voice. Using their focus and attention abilities they were able to get multiple information meaningfully in lesser time.

We find it suitable to further investigate this information design approach. We are presently in the process to develop a software that would be able to play multiple live programs simultaneously. Each program would have its own set of controls mapped with auditory perceptions. Users would be able to set the controls, i.e. they would be able to pan the stream, make the volume low and high, change the pitch, change the rate of voice streams and much more which may help them to listen to multiple voice streams simultaneously using their focus and attention abilities. This web-based software would be used to observe the interaction behaviour of users. For example, what values they set to the control to listen to the multiple sounds?

References

1. Li, G.-P., Huang, G.-Y.: The "core-periphery" pattern of the globalization of electronic commerce. In: Proceedings of the 7th International Conference on Electronic Commerce, ICEC'05, pp. 66–69. ACM, New York, NY, USA (2005)
2. Kazhamiakin, R., Bertoli, P., Paolucci, M., Pistore, M., Wagner, M.: Having services "your-way!": towards user-centric composition of mobile services. In Future Internet–FIS 2008, pp. 94–106. Springer (2009)
3. Kortum, Philip: HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2008)
4. Schnelle-Walka, D.: I tell you something. In: Proceedings of the 16th European Conference on Pattern Languages of Programs, p. 10. ACM (2012)
5. World Health Organization. Visual impairment and blindness. http://www.who.int/mediacentre/factsheets/fs282/en/ (2014). Accessed on 04 Jan 2016
6. Csapó, Ádám, Wersényi, György: Overview of auditory representations in human-machine interfaces. ACM Comput. Surv. (CSUR) **46**(2), 19 (2013)
7. Sato, D., Zhu, S., Kobayashi, M., Takagi, H., Asakawa, C.: Sasayaki: augmented voice web browsing experience. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'11, pp. 2769–2778. ACM, New York, NY, USA (2011)

8. Church, K., Cherubini, M., Oliver, N.: A large-scale study of daily information needs captured in situ. ACM Trans. Comput. -Hum. Interact. **21**(2), 10:1–10:46 (2014)
9. Agarwal, S.K., Jain, A., Kumar, A., Nanavati, A.A., Rajput, N.: The spoken web: a web for the underprivileged. SIGWEB Newsl. (Summer), 1:1–1:9 (2010)
10. Paule-Ruiz, M.P., Álvarez García, V., Pérez-Pérez, J.R., Riestra-González, M.: Voice interactive learning: a framework and evaluation. In: Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education, ITiCSE'13, pp. 34–39. ACM, New York, NY, USA (2013)
11. Patel, N., Chittamuru, D., Jain, A., Dave, P., Parikh, T.S.: Avaaj otalo: a field study of an interactive voice forum for small agriculturers in Rural India. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'10, pp. 733–742. ACM, New York, NY, USA (2010)
12. Lewis, J.R.: Practical Speech User Interface Design. CRC Press, Inc. (2010)
13. Schow, R.L., Seikel, J.A., Chermak, G.D., Berent, M.: Central auditory processes and test measuresasha 1996 revisited. Am. J. Audiol. **9**(2), 63–68 (2000)
14. Canosa, R.L.: Real-world vision: selective perception and task. ACM Trans. Appl. Percept. (TAP) **6**(2), 11 (2009)
15. Guerreiro, J.: Using simultaneous audio sources to speed-up blind people's web scanning. In: Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, p. 8. ACM (2013)
16. Hussain, I., Chen, L., Mirza, H.T., Chen, G., Hassan, S.-U.: Right mix of speech and non-speech: hybrid auditory feedback in mobility assistance of the visually impaired. Univ. Access Inf. Soc. 1–10 (2014)
17. Guerreiro, J., Gonçalves, D.: Text-to-speeches: evaluating the perception of concurrent speech by blind people. In: Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility, pp. 169–176. ACM (2014)
18. Scott, B., Neil, T.: Designing Web Interfaces: Principles and Patterns for Rich Interactions. O'Reilly Media, Inc. (2009)
19. Dix, A., Finlay, J.E., Abowd, G.D., Beale, R.: Human-Computer Interaction (2003)

Separability Assessment of Selected Types of Vehicle-Associated Noise

Adam Kurowski, Karolina Marciniuk and Bożena Kostek

Abstract Music Information Retrieval (MIR) area as well as development of speech and environmental information recognition techniques brought various tools intended for recognizing low-level features of acoustic signals based on a set of calculated parameters. In this study, the MIRtoolbox MATLAB tool, designed for music parameter extraction, is used to obtain a vector of parameters to check whether they are suitable for separation of selected types of vehicle-associated noise, i.e.: car, truck and motorcycle. Then, cross-correlation between pairs of parameters is calculated. Parameters for which absolute value of cross-correlation factor is below a selected threshold, are chosen for further analysis. Subsequently, pairs of parameters found in the previous step are analyzed as a graph of low-correlated parameters with the use of the Bron-Kerbosch algorithm. Graph is checked for existence of cliques of parameters linked in all-to-all manner related to their low correlation. The largest clique of low-correlated parameters is then tested for suitability for separation into three vehicle noise classes. Behrens-Fisher statistic is used for this purpose. Results are visualized in the form of 2D and 3D scatter plots.

Keywords Vehicle-associated noise · Low-level features · Bron-Kerbosch algorithm · MIRtoolbox

A. Kurowski (✉) · K. Marciniuk
Faculty of Electronics Telecommunications and Informatics Multimedia Systems Department, Gdansk University of Technology,
Narutowicza 11/12, 80-233 Gdańsk, Poland
e-mail: adakurow@sound.eti.pg.gda.pl

K. Marciniuk
e-mail: karmarci@sound.eti.pg.gda.pl

B. Kostek
Gdansk University of Technology Faculty of Electronics, Telecommunications and Informatics Audio Acoustics Laboratory,
Narutowicza 11/12, 80-233 Gdańsk, Poland
e-mail: bokostek@audioacoustics.org

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_10

1 Introduction

Nowadays the legal requirements (European Union Journal 2014/345/EC) and social actions are forcing road administration into building more advanced and accurate traffic control systems [1]. The aim of these actions is to improve quality and safety of road transportation and communication systems. In the 9th Road Safety Performance Index Report, that summarizes, among others, statistics of deaths in road accidents per million inhabitants in 2014, Poland was ranked in the forefront of the Member States and candidate countries to the European Union (over 80 deaths per million inhabitants). Higher statistics are only in Lithuania, Bulgaria, Romania and Latvia. The first and the foremost cause of such statistics is excessive speed. Another source of accidents, however indirect, leading to death or bodily injury is road congestion, changing individual driver's behavior, which may include aggressive lane changing, U-turns, overtaking other vehicles after the traffic jam is over, etc. Also, it should be pointed out that self-driving car technology emerges, that gets a lot of attention nowadays, which alongside legal and social requirements aim at making cars safer [2]. However, all these need a reliable traffic density measurement and intelligent assessment of dynamic characteristics of road traffic.

A typical Intelligent Transport Systems (ITS) are build up from a network of sensors. The more accurate systems are usually consist in integration of several technologies that complement each other. The inductive loop vehicle detection system is the most common and accurate technique that provides basic traffic parameters such as volume, presence, speed, type and gap between two vehicles. They are built in the road surface in close proximity to the carriageway. Usually they are mounted in pairs per each line of the road. The biggest disadvantage is the need to embed them in the asphalt during the installation and repair that requires the closure of the whole line or even the roadway. Another problem is related with a proper recognition of a large truck that sometimes can be marked as a two or more individual vehicles. The second type of detectors rely on the video image processing (VIP). One camera can supervise more than one line of the road. They are easy to mount and provide a lot of useful data. The technology can be employed to tracking an individual car in the road network. The disadvantage of such an approach is that the functionality of the automatic vehicle identification (AVI) is highly connected with the lighting conditions and weather changes [1]. However, lately high performance cameras appear that are immune to even extreme weather conditions enabling reliable identification in all weather conditions.

The effectiveness of both types of systems decreases with an increase of speed of vehicles, this is why there is a need for additional sensors placed in the proximity of the road to improve the efficiency and reliability of such systems. Therefore, the main goal of this paper is to propose an experimental setup that consists in acquiring acoustic signals from acoustic sensors (microphones), extracting on that basis a vector of parameters which will be suitable for acoustic identification of a vehicle type, and finally performing automatic recognition of a vehicle type. First,

some acoustic recordings are made to prepare the input data. These input data will serve later as a core of the database which will be needed for decision-based classification. However, in these preliminary research study they serve for parameter extracting and analyzing them in the context of their separability. Subsequently, using Bron-Kerbosch, a seminal maximal clique enumeration algorithm, a search for cliques of parameters, linked in all-to-all manner, related to their low correlation is performed. The largest clique of low-correlated parameters is tested in the context of separating individual vehicle noise into classes. This is based on Behrens-Fisher statistic. Results are visualized in the form of 2D and 3S scatter plots. Finally, conclusions and future plans of this research study are outlined.

2 Preparation of the Input Data

Three groups of vehicles were selected for this study: cars, trucks and motorcycles which differ in size and construction as they have a direct impact on vehicle speed and associated noise. The goal was to investigate if these differences influence the emitted noise to the extent that this enables to classify a particular vehicle into separate types. Samples of the vehicle noise were collected from two sources: recordings of vehicle noise signals hosted in the Internet and published under the Creative Commons license and recordings which were made by the authors, the latter performed in controlled conditions. Audio signals were edited to obtain a quasi-stationary signal discerning between the moment of a vehicle passing through the point closest to the microphone. All excerpts were then normalized.

It was assumed that each signal sample in the created database should contain noise signal related to a single vehicle. This encountered a considerable difficulty in measurements due to the nature of the traffic stream in cities. That's why for the purpose of experiments, two separate recordings were conducted. The first one took place in a residential area, where the speed limit is 40 km/h (with car density of approximately 50 vehicles/hour). It occurred that such an experimental layout makes possible to discern between sound events. The observation time was 20 min which contained 10 samples (including motorcycles). The second location recordings were made on a national route which has a speed limit of 70 km/h and typical traffic density is over 500 vehicles/hour (3 % of heavy vehicles). In this case, due to high traffic volume and traffic lights at distances of 200 m, vehicle overlapping was high in percentage, so we managed to obtain 15 samples from the 20 min recording that contain only heavy and light cars. Recording situations are depicted in Fig. 1.

Due to various sources of audio signals and differences of audio formats, all samples were converted to WAVE audio format. For stereo samples only one channel was taken to analysis. The database of audio signals consists of 60 samples. It was split into three groups of 20 samples. Each group is associated with one of the considered vehicle classes. All audio files accessed through the Internet had a sample rate of 44100 Hz and higher, thus they are of good quality.

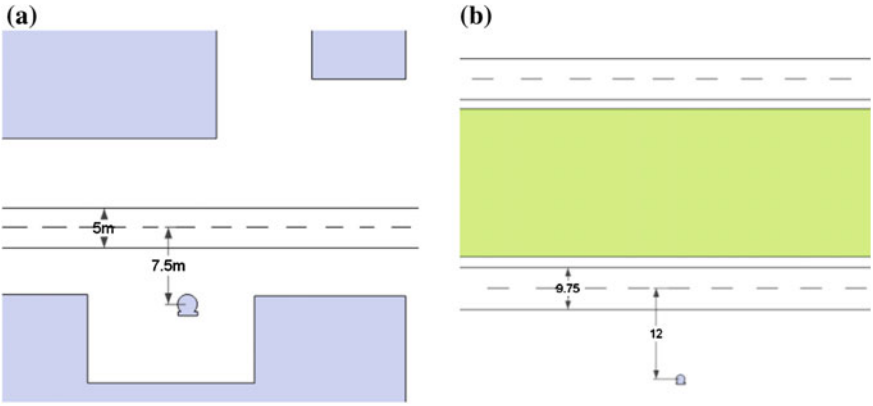


Fig. 1 Schematic presentation of the recording situations: in a residential area (**a**) and near a national route (**b**)

3 Analyses

In both time and frequency domains, sound recorded during passage of a vehicle depends on some external factors, as well as those associated with vehicle construction. The most important factors are the distance of the sensor from the source and the speed of the vehicle. In the lower gear noise generated by the engine (depending on the type, number of cylinders, capacity, etc.) plays greater role, contrarily for higher speed much more important is noise generated by tires and vehicle aerodynamic. When analyzing noise generated by vehicles, some researchers focus on the time domain [3, 4, 5], others look at frequency domain [6, 7, 8]. In most cases, however, analyses and parameterization are performed in both domains [9]. In this study, the last mentioned approach to vehicle signal parametrization was taken into account, as the goal of this study was to find uncorrelated parameters that will assure sufficiently high vehicle type recognition in the next stages of the study.

A set of 48 parameters was chosen for the purpose of the assessment of separation between classes. Parameter extracting was performed employing the MIRtoolbox package, which is typically used for tasks related to the music information retrieval (MIR) discipline, such as e.g. classification of music genres, evaluation of music emotions and others. Overall, it provides a number of algorithms for the purpose of the feature extraction from audio files which can also be used to extract data for classification of other types of audio signals, e.g. associated with noise generated by vehicle passage.

Temporal parameters taken into consideration were RMS power of the signal and zero crossing rate. Statistical measures of the spectrum such as centroid, skewness or kurtosis were also included. Another group of investigated parameters was timbre-related parameters which were implemented in MIRtoolbox [10]. Examples of such parameters are roughness or the harmonic change detection

function (HCDF). The last set of parameters used was vector of 31 mel-frequency cepstral coefficients (MFCC) parameters (including 0th coefficient related to average energy of the signal), which are commonly contained in construction of feature vectors and the classification systems [11, 12, 13]. A list of parameters apart from MFCC coefficients is shown in Table 1. A detailed description of each parameter calculated by the MIRtoolbox is provided in the manual of MIRtoolbox [10].

The analysis consisted of few steps which are shown in Fig. 2. The first step was extraction of the feature vector for each audio file from the database. The second step was calculation of cross-correlation matrix which contains correlation coefficients for each possible pair of parameters. As mentioned before, the goal of this research is to identify low-correlated features to be used in further analysis. Therefore, only pairs of parameters with cross-correlation value between them, smaller than a specified threshold, were taken into consideration. A threshold value was experimentally selected and was equal to 0.3. Such a list of pairs of low-correlated parameters may be considered as a list of a graph edges. Vertices of a graph are parameters and only low-correlated parameters are connected by the edges of the graph. For the next step of calculation, a structure of parameters connected with all-to-all manner was needed. Such a structure is called a clique [14]. Illustration of an example of such a graph containing clique of low-correlated parameters is depicted in Fig. 3.

The Bron-Kerbosch algorithm for maximum clique finding was employed [15]. The next step was to check if a set of parameters obtained from the clique finding algorithm could be used for separation of vehicle noise classes in a straightforward way, without employing a decision system. In order to evaluate this possibility, the Behrens-Fisher statistic was calculated for each parameter and pair of classes. It is given by the following formula:

$$V = \frac{\overline{P_X} - \overline{P_Y}}{\sqrt{\frac{S_{PX}}{n_X} + \frac{S_{PY}}{n_Y}}}, \tag{3.1}$$

Table 1 Parameters used for construction of the feature vector in addition to MFCC parameters

Temporal parameters	Spectral parameters	Timbre and dynamics parameters
RMS power of audio sample	Spectral centroid	Mean roughness
Zero-crossing rate	Spectral skewness	Standard deviation of roughness
	Spectral kurtosis	Entropy of roughness
	Spectral flatness	Mean HCDF
	Spectral entropy	Standard deviation of HCDF
		Entropy of HCDF
		Mode
		Roll-off
		Brightness
		Low energy

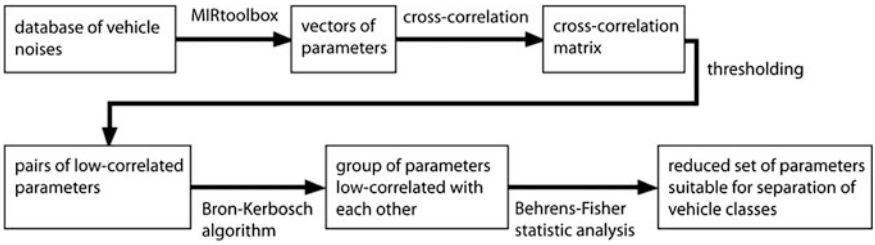


Fig. 2 Diagram of the algorithm implemented for the purpose of finding feature vector

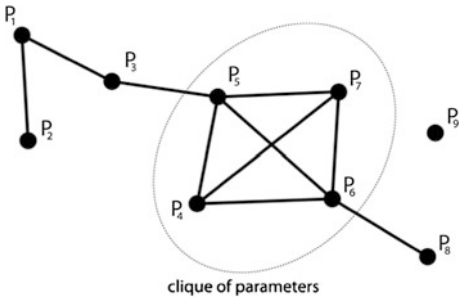


Fig. 3 Graph of low-correlated parameters and a clique of such a kind of parameters

where n_X denotes number of samples related to class X, $\overline{P_X}$ is mean value of parameter P_X and S_{PX} is estimator of variance of this parameter. Only parameters which are associated with the absolute value of the statistic which is greater than a specified critical value $V_c = 2.086$ were used in further analysis [16]. This is the critical value associated with sets containing 20 samples. The last step of this study was creating scatter plots which enables to observe data set separation. Both two-dimensional and three-dimensional visualizations were prepared. These visualizations are shown in the next Section.

4 Results

Three series of comparisons of pairs of vehicle noise classes were performed. For each comparison, a set of parameters best separating those classes was calculated. Resulting values of the Behrens-Fisher statistic for each case are shown in Table 2. Four of parameters shown in Table 2 exceeded critical value of the Behrens-Fisher statistic only for one pair of vehicle noise classes and may be used only for separation in this single case. Other four parameters were suitable for separation of two pairs of vehicle noise classes and the last two were found to be universal for all three pairs of classes. In addition to classification based on MFCC parameters, it is also possible to obtain separation with the use of parameters related to the timbre and dynamics implemented in MIRtoolbox. Moreover, these parameters provide the

Table 2 Values of Behrens-Fisher statistic of best separating parameters for each pair of considered vehicle noise classes

Parameter name	Car/Truck	Car/Motorcycle	Truck/Motorcycle
Mean roughness	3.28	2.33	*−0.36*
HCDF standard deviation	*−1.51*	−4.61	−3.27
HCDF entropy	*−0.29*	2.50	2.70
0th MFCC	−2.75	*0.33*	3.01
2nd MFCC	−2.87	−5.78	−2.63
3rd MFCC	−2.46	−5.56	−2.46
8th MFCC	2.65	*1.91*	*−0.82*
17th MFCC	*−0.50*	*1.95*	2.16
27th MFCC	−2.13	*−1.13*	*0.23*
31th MFCC	*0.67*	2.29	*1.49*

best value of Behrens-Fisher statistics when separation of car/truck and truck/motorcycle classes are considered. An example of such separation in the two-dimensional space of parameters is shown in Fig. 4. A three-dimensional case is depicted in Fig. 5.

Full separation of classes was not possible while employing the extracted parameters, however it was possible to discern areas of high concentration of entities belonging to the one class of vehicles. Such careful parameter extraction may be useful for designing a machine learning solution for the purpose of vehicle sound classification. It is also worth mentioning, that parameters, which are typically used for music information retrieval were found to be useful for classification of the vehicle noise type.

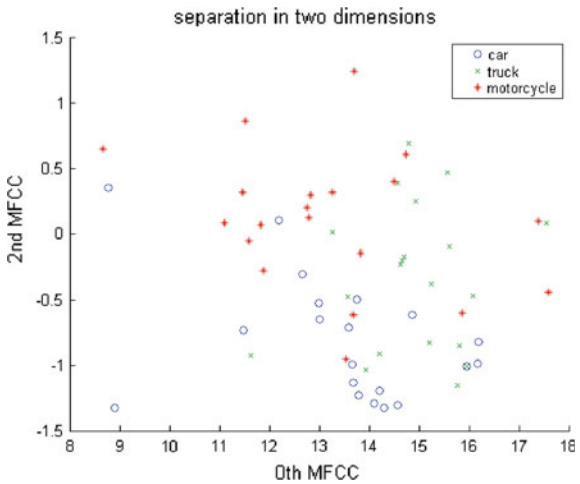


Fig. 4 Scatter plot illustrating separation of parameters in the two-dimensional parameter space

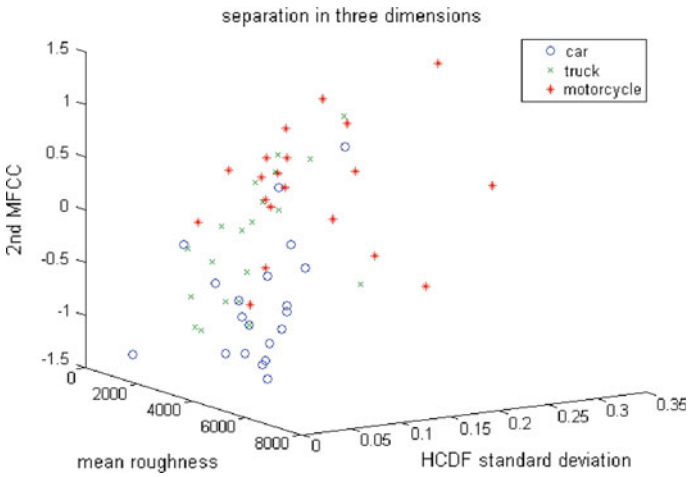


Fig. 5 Scatter plot illustrating separation of parameters in the three-dimensional parameter space

5 Conclusions

The approach presented in this work may be used as a starting point to build a decision system for recognition of a vehicle type on the basis of signal captured by microphones. Parameters derived from the analysis performed, based on both time- and frequency domains, showed their usability in separation vehicle noise sources into investigated classes. It is worth mentioning that specific descriptors related to timbre and dynamics of a signal were also found to be useful.

Tools developed for the purpose of MIR are widely used and therefore tested by many users. They also offer user-friendly set of methods intended for processing acoustic signals. They may be further extended by automating the process of selection of uncorrelated parameters and choice parameters which will then provide the best separation of considered classes of signals. Automation of mentioned tasks may be obtained by using graph theory algorithms such as Bron-Kerbosch and analysis of statistical measures employing Behrens-Fisher statistic.

Acknowledgments This research was supported by the Polish National Centre for Research and Development within the grant No. OT4-4B/AGH-PG-WSTKT.

References

1. Klein, L, Mills, M., Gibson, D.: Traffic Detector Handbook. 3rd edn., vol. I. U.S. Deparment of Transportation, Federal Highway Administration (2006)
2. Rajasekhar, M., Jaswal, A.: Autonomous vehicles: the future of automobiles. In: 2015 IEEE International Transportation Electrification Conference (ITEC), Chennai, India (2015)

3. Paulraj, M., Adom, A., Sundararaja, S., Rahima, N.: Moving vehicle recognition and classification based on time domain approach. In: Malaysian Technical Universities Conference on Engineering & Technology 2012, Kangar Perlis, Malaysia (2012)
4. Sampan, S.: Neural fuzzy techniques in vehicle acoustic signal classification. Ph.D. dissertation, Department of Electrical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA (1997)
5. Chen, S., Sun, Z., Bridge, B.: Traffic monitoring using digital sound field mapping. IEEE Trans. Veh. Technol. **50**, 1582–1589 (2001). doi:10.1109/25.966587
6. Huadong, W., Siegel, M., Khosla, P.: Vehicle sound signature recognition by frequency vector principal component analysis. In: IEEE Instrumentation and Measurement Technology Conference, St. Paul, Minnesota, USA (1998)
7. Wellman, M., Srour, N., Hills, D.: Acoustic feature extraction for a neural network classifier. Army Research Laboratory. Technical report ARL-TR-1166, Army Research Laboratory (1997)
8. Li, D., Wong, D., Sayeed, A.: Detection, classification and tracking of targets in distributed sensor networks. IEEE Signal Process. Mag. **19**, 17–29 (2002). doi:10.1109/79.985674
9. Borkar, P., Malik, L.: Review on vehicular speed, density estimation and classification using acoustic signal. Int. J. Traffic Transp. Eng. **3**, 331–343 (2013). doi:10.7708/ijtte.2013.3(3).08
10. MIRtoolbox 1.5 Users Manual. https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/MIRtoolbox1.5Guide. Accessed on 14 Apr 2016
11. Sivakumar, S., Gavya, P.: Location estimation in wireless sensor network using H-PSO algorithm. In: IJCA Proceedings on International Conference on Innovations in Computing Techniques, Karachi, Pakistan (2015)
12. Kazi, F., Bhalke, D.: Musical instrument classification using higher order spectra and MFCC. In: 2015 International Conference on Pervasive Computing (ICPC), Pune, India (2015)
13. Paulraj, M., Yaacob, S., Nazri, A., Kumar, S.: Classification of vowel sounds using MFCC and feed forward neural network. In: 5th International Colloquium on Signal Processing & Its Applications, Kuala Lumpur, Malaysia (2009)
14. Bondy, J., Murty, U.: Graph Theory with Applications. Elsevier Science Ltd., Oxford (1976)
15. Cazals, F., Karande, C.: Note on the problem of reporting maximal cliques. Theoret. Comput. Sci. **407**, 564–568 (2008)
16. Mason, R., Gunst, R., Hess, J.: Statistical Design and Analysis of Experiments: With Applications to Engineering and Science. 2 edn. Wiley, Hoboken (2003)

Popular Brain Computer Interfaces for Game Mechanics Control

Dominik Szajerman, Michał Warycha, Arkadiusz Antonik and Adam Wojciechowski

Abstract Brain computer interfaces become more and more available, mainly due to cheaper and better technology. Some of the devices, like NeuroSky MindWave and Emotiv EPOC became an example of the affordable apparatus that may be exploited for game interaction. At the same time most of authors, exploring brain waves interactions paradigms in games, concentrate on professional EEG devices, equipped with even hundreds of sensors, assuring high quality encephalography measures, what obviously outperform technically simplified solutions. Thus the paper provides an analysis of MindWave and Emotiv applicability for selected game interaction tasks: moving the object, selecting one of a few possibilities and interaction with the help of dialogue system. A corresponding game environment experiments were performed and analysis of control paradigms was provided.

Keywords Control paradigms · Brain waves · EEG · Computer games

1 Introduction

Brain computer interfaces (BCI) can capture and interpret brain activity in a form of electrical signals detected on scalp, cortical surface or within the brain. Electrodes assigned to the scalp should record signal that is degraded by covering of the brain tissue (i.e. skull, scalp) and just a synchronized activity of a large numbers of neural elements can be detected [1]. Unfortunately electrodes register also activities other than the brain, like electrical power lines noise or biological noise from muscles, heart or eyes. These are the main aspects limiting frequency of the brain monitoring.

As BCI may not depend on neuromuscular control it can provide alternative mean of communication. It might be useful not only for people with neuromuscular disorders or motor impairments but for casual users as well. Except professional med-

D. Szajerman (✉) · M. Warycha · A. Antonik · A. Wojciechowski
Institute of Information Technology, Łódź University of Technology, Łódź, Poland
e-mail: dominik.szajerman@p.lodz.pl
URL: http://it.p.lodz.pl/

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_11

ical electroencephalography (EEG) (i.e. g-tec solutions) or functional magnetic resonance (fMRI) apparatus, several less professional and at the same time cheaper devices are available. Among them, game dedicated devices, aiming to popularize brain waves analysis for interface control were proposed (i.e. Emotiv EPOC, NeuroSky MindWave, etc.).

Unfortunately in current literature more attention is put to professional, less available, EEG solutions as they reveal better quality of signal [2]. Our contribution was Emotiv and Mindwave performance analysis in the context of three popular interface actions, which are popular in current games interaction activities: a moving object steering, selecting one of a few possibilities and extension of interaction possibilities with the help of dialogue system.

2 Related Work

Electroencephalography (EEG) signal acquisition is performed by an array of electrodes attached to user scalp—usually in a form of a cap or dry electrodes hidden in a nicely designed head mounted device. Signal analysis can be performed synchronically, then a correspondence with expected brain activity is necessary, or no synchronically but then holistic signal analysis and classification should be performed [1]. Collected signals should be then classified with respect to the noises and biases coming from the environment or human.

Predominant approach to brain waves analysis bases on frequency and amplitude. In this context subsequent frequency bands (rhythms) can be distinguished: gamma (above 40 Hz), beta (≈12–≈28 Hz), alpha (≈8–≈13 Hz), mu (≈8–≈12 Hz), theta (≈4–≈7 Hz) and delta (≈0.5–≈3 Hz). That is why in brain computer interfaces several potential bands, evoked by process of thinking or concentration, can be exploited for human brain activity analysis.

BCI application for games was analysed within several studies evaluating researchers, developers and users points of view [2]. Among selectively emphasised aspects, considered as brake through for BCI games, these are easiness of playing and new development platform. Some authors [3, 4] even suggest that non-medical BCI will be soon the main brain computer interfaces application.

Comparing to traditionally known game interfaces, games exploiting brain activity analysis appeared to be not efficient in speed and accuracy performance [5]. It was perceived to be the main bottleneck of successful BCI games commercialisation. Though Vidal [6] reported for the visually evoked potential (VEP) the information transfer rate (ITR) of about 170 bits per minute, state of art reviews [7] claim to achieve information transfer rate of about 10–25 bits per minute. It is really incomparable with traditional WIMP interfaces which information transfer rate is measured in millions bits per minute—sensitivity of human senses and bandwidth of corresponding sensors is considerably higher [8].

Still not perfect characteristics of BCI interfaces are attempted to be strenghten by its conscious correlating with game design rules. Gurkok proposed two main descrip-

tors that should be considered in BCI games context [9]. These are correspondence between psychological needs and game playing motivation as well as correspondence between action and reaction of the user to events performed in BCI application. Whereas interaction, as the most latency critical, seems to be more important.

Thus BCI games experiments focus rather on surveying BCI potential rather then regular gameplay development. Experiments in game area [5, 10, 11] exploit mainly professional EEG devices like Biosemi Neuroscan [10]. Several old-fashioned games like Pacman, Pinball and others, controlled with brain waves have been already published [12–15]. Among them several control paradigms were analysed. Hjelm [16] measured relaxation score, derived from alpha and beta rhythms ratio for controlling the ball movement across the table, whereas metaphor of Pinball was also proposed by Tangerman [13]. Pires [14] considered event related potential (P300 ERP) for Tetris play within attention-deficit and hyperactive disorder (ADHD) children. Experiments revealed that users usually failed while position selection but specifying of the objects (Tetris blocks) moving direction has appeared to be effective. Pineda [17] proposed a first person shooter game control by means of mu rhythm power control. Krepki [18] and Reuderink [12] proposed a BCI PacMan implementation. In this experiment user's avatar made each step every 2 s but users sometimes reported PacMan proper movement before conscious decision was made by the user. This suggests new level of interaction, only available for BCI.

Nevertheless, low cost and less professional brain computer interfaces are scarcely discussed in a context of gameplay mechanics development. Presented paper main contribution is to reveal popular, home use designed devices characteristics and their application and evaluation for selected computer games mechanics.

3 Data Acquisition and Processing

There are two EEG devices considered in this work: NeuroSky MindWave and Emotiv EPOC (Fig. 1). Both are designed to popular use with personal computers. They bring the power of control the computer applications with thoughts—electrical activity of the brain. Both NeuroSky MindWave and Emotiv EPOC devices are described in terms of hardware and software design as well as implementation issues.

3.1 NeuroSky MindWave

NeuroSky MindWave consists of a headset, an ear-clip (including ground and reference electrodes), and an arm with the sensor (EEG electrode) which is resting on the forehead above eyes [19]. The device output data is divided into following groups:

- data related to signal quality—*poor_signal*, integer in the range 0–255, represents signal strength and noise, zero means the best quality,

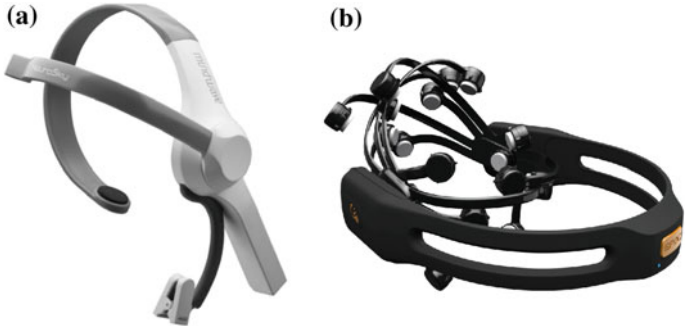


Fig. 1 **a** Neurosky Mindwave [19]. **b** Emotiv EPOC [20]

Fig. 2 Arrows presented to a user

- data concluded by the device internal software—levels of *attention* and *meditation*, normalised to interval [0–100],
- eye blinking (*blinkStrength*), the higher the value (1–255) the more closed the eye, very useful value, because it can be initiated intentionally in most cases,
- raw brain waves data are a few types (frequencies) of brain waves: delta, theta, low alpha, high alpha, low beta, high beta, low gamma and high gamma. This mode allows to develop and test own brain wave analysis algorithms.

As to elaborate raw brain wave values interpretation heuristics, the users were stimulated with two different pictures (Fig. 2): left arrow with green letter "L" and right arrow with red letter "R" alternately. Distinguished stimuli affect different brain parts. The colour centre in the ventral occipital lobe is responsible for colour vision [21], the parietal lobe allows to read (letters) [22], and hippocampus helps with directions recognition [23]. Such pictures were chosen in order to engage as many brain parts as possible and thereby achieve a maximum distinguishability of corresponding device responses. Resulting waves frequencies were analysed as to model the paradigm of thinking process.

The analysis covered 4 signals: raw brain waves, eye blinking strength, *attention* and *meditation* levels and gamma waves induced by body parts movement.

Raw brain waves were subjected to simple filtering and a back-propagation neural network with sigmoid activation function. Achieved results did not allow to build any accurate binary indicator, which could be used to control a game interface. Figure 3a shows an example of similar characteristics of delta waves independently on arrow image presented to a user.

Two signals given by the device in preprocessed form: *attention* (Fig. 3b) and *meditation* cannot be easily and directly controlled by a user. While the *attention* is somewhat controllable—a user can focus on something causing the increase of the *attention* level, the experiments showed, that *meditation* cannot be or is very difficult

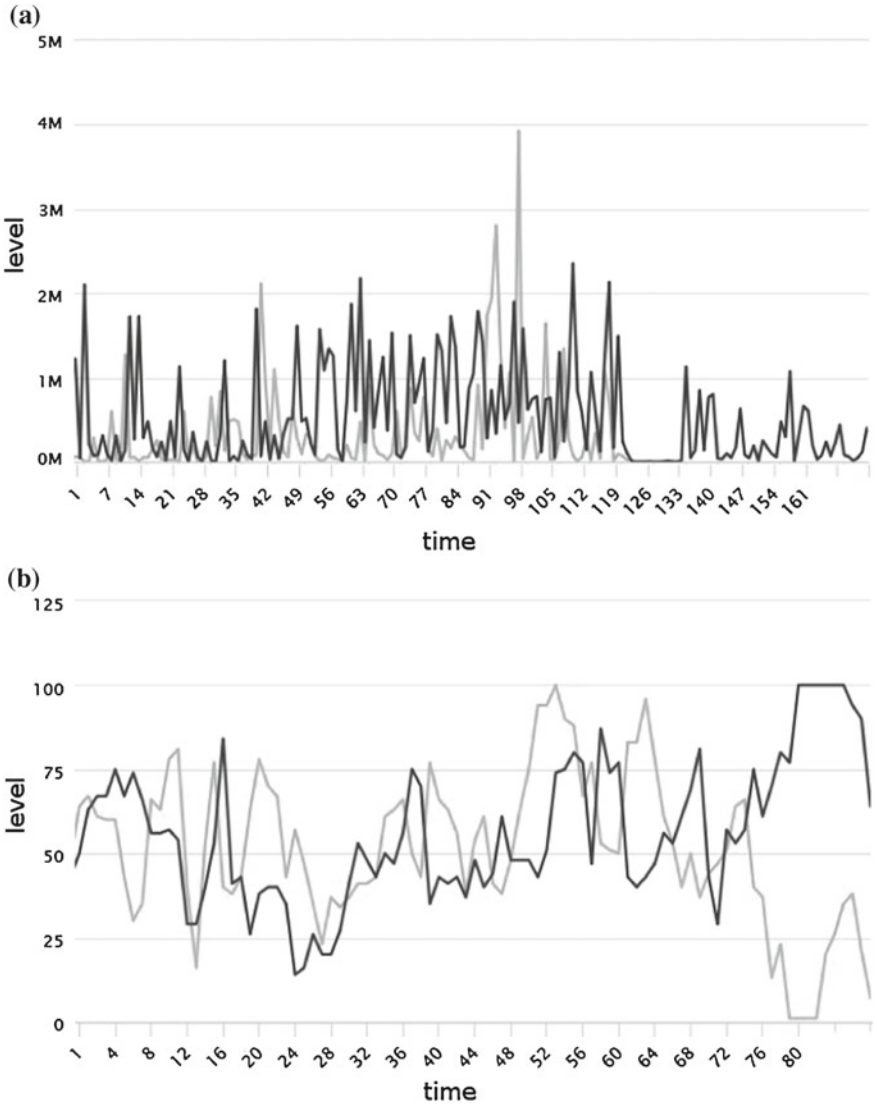


Fig. 3 **a** Delta brain wave insensitive to type of arrow image presented to a user (*black–left arrow*, *grey–right arrow*). **b** *Attention* level

to be changed by the user. Therefore the *attention* was only considered as potentially useful in interactive application controlling.

Eye blinking signal revealed utterly different characteristics. It can be fully controlled by a user and therefore can be simply thresholded in order to obtain useful, binary parameter (Fig. 4a).

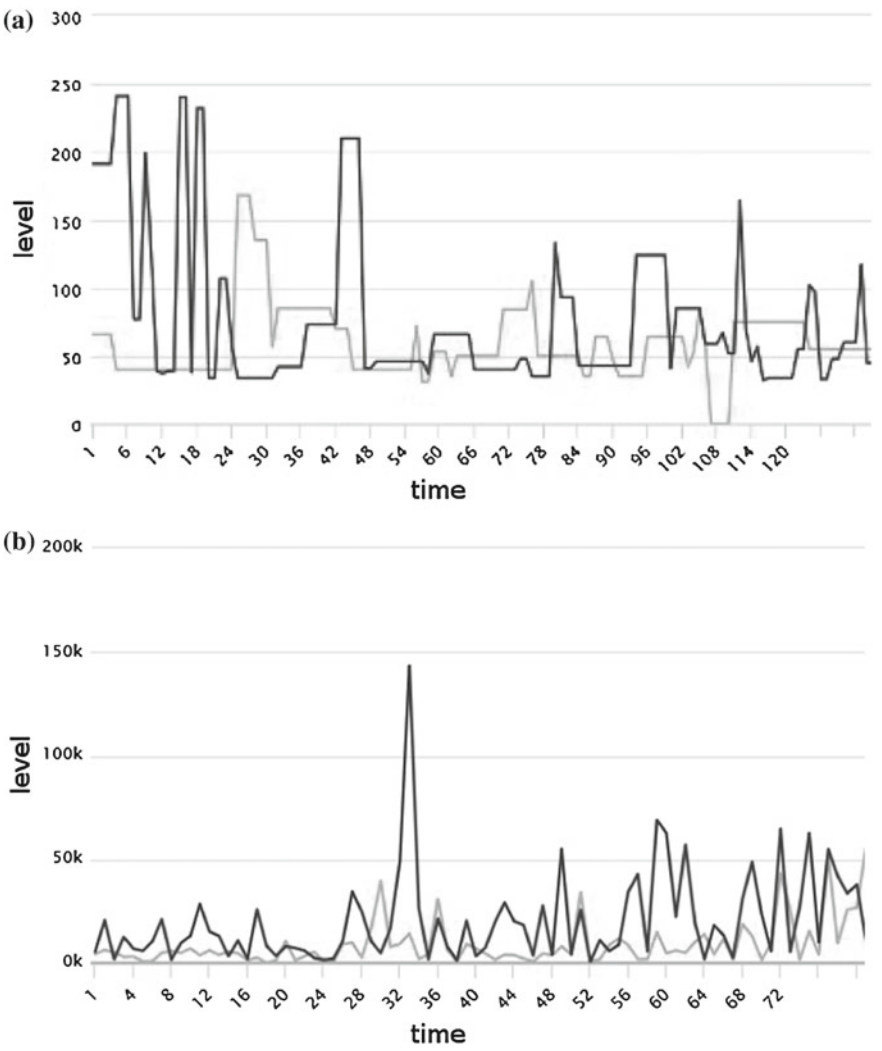


Fig. 4 **a** Recorded blinking strength. The *black line* indicates closing of an eye. The *grey line* indicates an open eye. **b** Recorded gamma wave. The values representing the movement of select body parts (*black lines*) are on average larger than these representing the immobility (*grey lines*)

During the analysis of *gamma* waves (Fig. 4b), induced by body parts movement, the knowledge, that movement of selected body parts, like tongue, fingers or toes increases the activity of gamma waves [24] proved useful. It became clear, that even with simple filtering and a back-propagation neural network with sigmoid activation function it was possible to clearly distinguish between the user's body parts movement and immobility states. Thus gamma brain waves were taken into account and it became possible to use them in order to control the game.

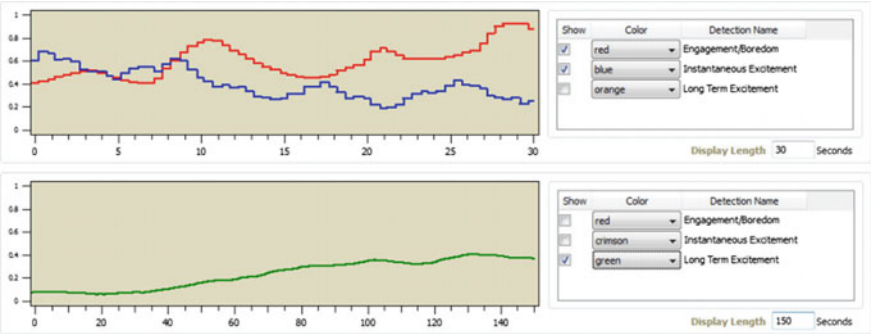


Fig. 5 Example preprocessed and normalized values acquired from Emotiv EPOC. Screen from Control Panel application [20]

3.2 Emotiv EPOC

Emotiv EPOC is a device capturing an electric field produced by the brain during its activity [20]. The device consists of 16 electrodes (14 EEG and 2 references), which measure the difference of the electric field over time. The device requires using a liquid to improve electrical contact of electrodes with the head's skin. It can be used in three main modes [20].

1. Cognitive Suite (aka Mental Commands) interprets the thoughts and intentions of the user. It can be understood as a kind of mental commands. Each user profile can contain the data for up to 15 different commands.
2. Expressive Suite (aka Facial Expressions) obtains the facial expression of the user. 8 EEG sensors can pick up signals from facial muscles and the eyes.
3. Affective Suite (Performance Metrics) detects level of 4 emotions: *excitement*, *engagement/boredom*, *frustration* and *meditation*.

The main advantage is that these three modes give preprocessed, normalized data (Fig. 5), which is easy to interpret. Furthermore, all data sources—modes can be accessible at the same time.

It was not necessary to prepare complex data interpretation algorithms or filtering. Simple thresholding of selected signals was enough, to achieve desired reaction. The testing applications could be developed directly with the help of the producer's software.

4 Tests

In order to prove the effectiveness of the interaction with the help of presented methods three applications were built. They test six different data sources on both devices—attention level, gamma waves activity and eye blinking on NeuroSky MindWave and Cognitiv, Expressiv and Affectiv modes on Emotiv EPOC.

The three control paradigms were tested:

1. The control of an object with an unlimited freedom of movement in two-dimensional space. Considering this freedom, in order to give the user a feeling of complete control, there were used the signals which could be initiated by the user intentionally. It was ensured by the methods developed for MindWave.
2. The control of game objects, as a choice of range from 1 to 4 possible moves. It is decision making from finite and small set of possibilities. It is good example for testing of mental commands provided by EPOC.
3. The control of a game engine dialogue system. It allowed to test the possibility of change a well-known games standard, to extend it by quite new mechanisms. The dialogue trees were expanded by quite new paths. The paths revealed in specific circumstances related to the user's behaviour. It was done with EPOC and it emotions and facial expressions modes.

The tests were performed on a grup o 7 people: 5 men and 2 women aged between 21 and 24.

4.1 NeuroSky MindWave-Based Eye Blinking Detection

The purpose of the first application was to make use of the acquired data to build a real-time control paradigm. The controlled object was a ball rolling on the floor (Fig. 6). The ball could have its movement direction changed according to two input signals. The former was the attention level and it controlled the velocity of the ball. The latter was the eye blinking activity which allowed the user to turn left or right. Every blink rate caused turning direction changes.

The test showed that the eye blinking strength gives the user a satisfactory level of control over the ball movement, because it was possible to achieve the target in every attempt.

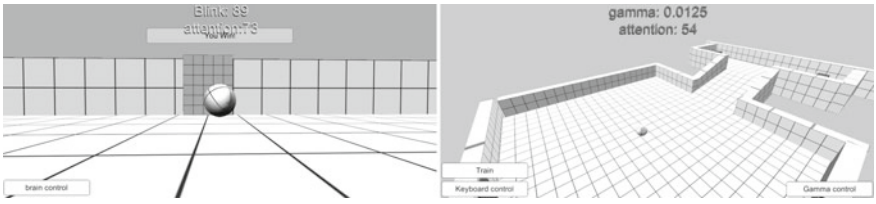


Fig. 6 The ball controlled with the attention level and eye blinking level (*left*) and gamma wave activity level (*right*). These levels are shown in the top of the pictures

4.2 NeuroSky MindWave-Based Gamma Wave Activity Detection

The same application implemented another NeuroSky MindWave control method—with the help of gamma wave activity (Fig. 6). Basing on observation, that some body parts (such as tongue, fingers or toes) movement raises the gamma wave level, it was possible to adapt it to control the ball. The low gamma level (below 0.5) corresponded to the direction to the left, and the high one (above 0.5) to the right.

The test of gamma wave activity steering proved significant inaccuracy of the ball control. In most cases it was not possible to achieve the target and the user had to switch to keyboard aid.

4.3 Emotiv EPOC Cognitiv Suite Mode

The second application was designed to test mental commands in terms of recognition and distinguishability. The controlled object was a *Game of Fifteen* (Fig. 7). The four commands: lift (meaning up), drop (meaning down), left and right were detected during gameplay. After the slide conditions were checked the suitable tile was moved in the specified direction. This control method demanded additional processing of the incoming data. It was necessary to introduce some delays and condition checks in order to avoid the multiplication of commands which might be accidentally generated by the user before the signal returns to neutral (no command).

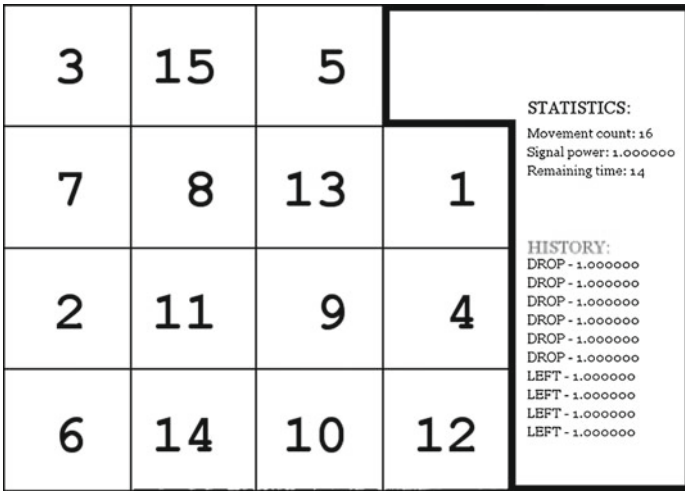


Fig. 7 Screen from Game of Fifteen controlled by mental commands. The current command is drop (*down*), the previous was left

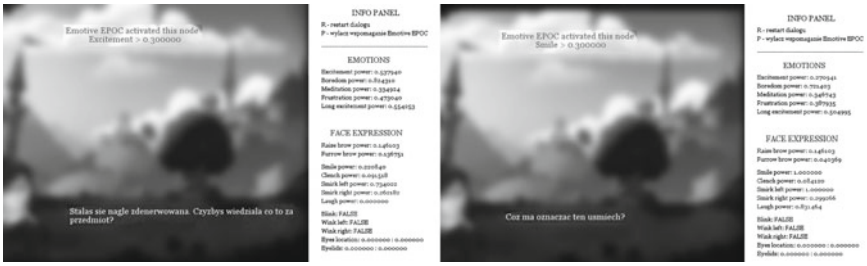


Fig. 8 *On the left* Affectiv mode detected excitement—it caused switching the dialogue path in order NPC to say: "You became nervous. Do you know what is this, don't you?". *On the right* Expressiv mode detected smile—it caused switching the dialogue path in order NPC to say: "What does this smile mean?"

The test involved placing the tiles in order and measuring of the time. It showed that the method gives the user a good level of control over the puzzle. It was possible to achieve significant speed of sliding—one slide in less than 3 s. Such speed could be considered interactive and comfortable.

4.4 Emotiv EPOC Expressiv and Affectiv Modes

The third application was based on two data sources: facial expressions and emotions. Both were used to control another type of interaction—conducting a dialogue. The controlled object was a computer game dialogue system. It has some standard features like the data structure of texts, questions, answers, conditions and decision making nodes. Furthermore, the system was expanded by a new type of *emotive nodes* which could switch between different dialogue paths according to fulfilment of conditions concerning given levels of facial expression features or emotions. Figure 8 presents the parts of the sample gameplay, where using EPOC, the player managed to run the options first with emotions, and then using facial expressions.

The test presented a new approach of building dialogue systems for computer games. Thanks to it dialogues with non-player characters could be more sophisticated and give the user greater immersion.

5 Conclusions

NeuroSky Mindwave is convenient in use due to its construction. Furthermore there is no need to use additional substances supporting data reception. It provides three types of data: raw brain waves, brain metrics and eye blinking strength. This information is calculated by internal device's algorithms. Among the others, the eye blinking

and the gamma waves activity (in response to body part movement) proved to be the most accurate and suitable for use in interactive applications. The device gives the opportunity to use it as a controller assuming that a user would not have too much freedom of movement in order to minimise the noise. The signal quality can be also improved by silent environment which is free of electrical equipment as much as possible.

Whereas Emotive EPOC requires more maintenance (providing humidity of electrodes) than NeuroSky MindWave, from a end-user point of view, it gives better opportunities in terms of effective software building. It is much convenient to use processed data (commands, emotions and facial expressions), than to analyse raw brain waves. Unfortunately internal algorithms work better, when they are calibrated during relatively long process of training, which does not give the guarantee of success (100 % accuracy).

Apart from this, both devices proved to be suitable to moving the object with an unlimited freedom of movement—where the eye blinking strength proved to be more useful than gamma waves activity, selecting one of a few possibilities with the help of mental commands and interaction with the help of dialogue system enriched by emotions and facial expressions detection. All three control paradigms were adapted with satisfactory convenience and precision.

References

1. Kołodziej, M.: Przetwarzanie, analiza i klasyfikacja sygnału EEG na użytek interfejsu mózg—komputer. Ph.D. thesis, Politechnika Warszawska (2011)
2. Ahn, M., et al.: A review of brain-computer interface games and an opinion survey from researchers, developers and users. Sensors **14**(8), 14601–14633 (2014)
3. Future BNCI. A roadmap for future directions in brain/neuronal computer interaction (2012)
4. Van Erp, J., et al.: Brain-computer interfaces: beyond medical application. Computer **45**, 26–34 (2012)
5. Bos, D.P.-O., et al.: Brain-computer interfacing and games. In: BCI, pp. 149–178 (2010)
6. Vidal, J.J.: Real-time detection of brain events in EEG. Proc. IEEE **65**(5), 633–641 (1977)
7. Wolpaw, J.R., et al.: Brain-computer interfaces for communication and control. Clin. Neurophysiol. **113**(6), 767–791 (2002)
8. Strumiłło, P.: Personal navigation systems for the blind and visually impaired. Lodz University of Technology (2012)
9. Gürkök, H., Nijholt, A., Poel, M: Brain-computer interface games: towards a framework. In: Entertainment Computing—ICEC 2012. LNCS, pp. 373–380 (2012)
10. Martinez, P., Bakardjian, H., Cichocki, A.: Fully online multicommand brain—computer interface with visual neurofeedback using SSVEP paradigm. Comput. Intell. Neurosci. **13** (2007)
11. Gürkök, H., et al.: Evaluating a multi-player brain—computer interface game: challenge versus co-experience. Entertain. Comput. **4**(3), 195–203 (2013)
12. Reuderink, B., et al.: Affective Pacman: a frustrating game for brain-computer interface experiments. In: Intelligent Technologies for Interactive Entertainment. LNCS, pp. 221–227 (2009)
13. Tangermann, M., et al.: Playing pinball with non-invasive BCI. In: Advances in Neural Informat. Processing Systems, vol. 21, pp. 1641–1648. MIT Press, Cambridge (2009)
14. Pires, G., et al.: Playing Tetris with non-invasive BCI. In: Proceedings of the 2013 IEEE 2nd International Conference on Serious Games and Applications for Health (SeGAH), pp. 1–6 (2011)

15. Van de Laar, B., et al.: Experiencing BCI control in a popular computer game. IEEE Trans. Comput. Intell. AI Games **5**, 176–184 (2013)
16. Hjelm, S.I.: Research + design: the making of brainball. Interactions **10**(1), 26–34 (2003)
17. Pineda, J.A., et al.: Making a brain—computer interface possible. IEEE Trans. Neural Syst. Rehabil. Eng. **11**(2), 181–184 (2003)
18. Krepki, R., et al.: The Berlin brain - computer interface (BBCI)—towards a new communication channel for online control in gaming applications. MTA **33**(1), 73–90 (2007)
19. Brainwave Sensing Headset, NeuroSky. http://store.neurosky.com/pages/mindwave. Cited 13 Apr 2016
20. Emotiv EPOC. Emotiv, Inc. http://emotiv.com/epoc. Cited 13 Apr 2016
21. Winawer, J., Horiguchi, H., Sayres, R.A., Amano, K., Wandell, B.A.: Mapping hV4 and ventral occipital cortex: the venous eclipse. J. Vis. **10**(5) (2010)
22. Cohen, L., Dehaene, S., Chochon, F., Lehéricy, S., Naccache, L.: Language and calculation within the parietal lobe: a combined cognitive, anatomical and fMRI study. Neuropsychologia **38**(10), 1426–40 (2000)
23. Hillman, K.: A list of brain areas and what they do. Evolution. Psychol. (2014)
24. AlZu'bi, H.S., Al-Nuaimy, W., Al-Zubi, N.S.: Sixth International Conference on Developments in eSystems Engineering (DeSE), Abu Dhabi (2013)

Pervasive System for Determining the Safe Region Among Obstacles: Mobile Doctor on the Road Case Study

Hanen Faiez and Jalel Akaichi

Abstract Like other many fields, telemedicine has benefited from pervasive and ubiquitous access to knowledge granted by the internet and mobile wireless. Remote monitoring and diseases management are considered as most fasted growing areas within this field. Thanks to mobile communication technologies, humans today are able to provide patient with w better quality life in critical and emergency situations. In such a scenario, the patient needs to reach as soon as possible the health care provider and/or medical institution. To do so, he needs to go by a road without obstacles. In different research reviews, rooting algorithms that have been presented are treating the shortest path, the nearest path. In this paper, we present a new pervasive system able to find the Safe Area without mobile obstacles and impediments based on an incremental algorithm called Safe Region algorithm.

Keywords Pervasive healthcare · Emergency systems · Safe region · Mobile obstacles

1 Introduction

Humans today are able to provide patient with w better quality life in emergency situations. In such cases, the patient needs to reach as soon as possible the health care provider and/or medical institution. Same for the doctor who needs to be positioned and able to specify his destination in terms of suitability of this destination. In such emergency context, to quickly attain his destination he needs to go by a road without obstacles. In different research reviews, rooting algorithms that have been presented are treating the shortest path, the nearest path… etc. In this

H. Faiez (✉)
Bestmod, ISG Tunis Université de Tunis, Tunis, Tunisia
e-mail: hanenfaiez89@gmail.com

J. Akaichi
King Khalid University Guraiger, Abha, Saudi Arabia
e-mail: Jalel.akaichi@kku.edu.sa

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_12

paper, the conducted research is a routing algorithm aiming to increase routing efficiency by identifying the Safe Area in a road network. As already mentioned above, in front of an emergency situation the doctor must react quickly to save their patient's life, particularly if they are geographically dispersed, is then called to treat them remotely. He is therefore called to deliberate the road or to look for a road devoted from problems and obstacles, in a Safe Area. Our approach is a continuation of our previous work aimed to identify the most suitable route. We propose an algorithm that specifies the area without problems neither fixed nor mobile once detected due to prolifiration of objects on the network map. The remainder of this paper is structured as follows: we briefly outline the relevant background and related work in Sect. 2, followed by a description of the proposed system architecture, Sect. 3 describes the motivating scenario. Then, in Sect. 4 we present our experimental evaluation and Sect. 5 summarizes our results and concludes the paper.

2 Background and Related Work

As aforementioned, in emergency cases and when patients are geographically dispersed, the doctor must react in minimum of time to save them. He is therefore called to find them as quickly as possible suitable institutions for medical intervention. Several are factors that come into play in reducing the waiting time for the arrival of the emergency physician at the reception of an emergency call, one of these factors is how to guide the user to the appropriate road. We are talking about Route Calculator Algorithms. In fact many calculator algorithms have been proposed for the determination of nearest neighbors on the road network. In this section we will see in detail the principle of the most important route calculator approaches, techniques and algorithms found in the literature. In [1] the author proposed a new technique which aims to transform a road network into a higher dimensional space in order to utilize computationally simple metrics.

Four techniques for finding the first nearest neighbor to a moving query object on a given path were also proposed in [2].

Incremental network expansion solution (INE): The author in [3] presented a solution called INE for NN queries in spatial network databases. INE technique collects the space entities (e.g. medical institutions) of the network, and preserves their cartesian coordinates and connectivity between them. INE is based on an iterative process which takes as input a road network and generates a graph. Formally, this technique relies on the expansion of the network (based on a query point q) and examines in the graphs segments, points in the order in which they are met. The computation of distances is done by the application of the Dijkstra's algorithm.

Indeed, the algorithm starts by locating the segment containing q, then, it searchs all points of interest which are on the same segment. To say it in simplistic way, INE first locates the segment which contains the point query q, and checks all entities within it. The algorithm locates the segment that contains the query point q

and looks for all the entities within it. For example if the segment n_1 n_2 does not contain any entity and the query point q is closest to n1 than endpoint n_2, the closest node n_1 will be developed. While the n_2 point is placed in Q. Developing n_1, turns out that the end point is n_7, so the search is now in n1n7 and no entity is found within. N_7 is then stored in Q this time contains the following elements = <(n_2, 5), (N_7, 12)>. Since the distance to less than the distance to $n_2$est n_7, the point will be n_2. More details on this algorithm can be provided in [4]. The main disadvantage of this approach is that it is applicable only for the search of dispersed points in space.

Voronoi-based network nearest neighbor(VN3): Authors in [5] proposed a new approach for NN queries in spatial network databases, called VN3 (Voronoi-based network nearest neighbor). Their approach is mainly based on Voronoi diagrams properties. Adding to Voronoi diagrams principle, VN3 algorithm is based on so-called "localized pre-computation" of the network distances for a very small set of neighboring points in the network. The idea is to use directly the NVPs (network Voronoi polygon) of an NVD (network Voronoi diagram) to find the first nearest neighbor of a point query q. Thereafter, Voronoi polygons which are adjacent to the polygon in question can be used to determine a new set of candidate neighbors to the point query q.

Finally, the pre-computed distances are then used to redefine a new set of points as a function of the minimum distance between the query q and the candidate point c in the new set. That is to say, the filter process is iterative and must be invoked k times to find the first k nearest neighbors of q. VN3 consists of the following steps: Pre-calculation of the solution space, utilization of an index structure and pre-computation of the exact distances for a very small portion of data. Experiments with several real-world data sets showed that VN3 outperforms INE [2] by up to an order of magnitude. However, in the case the number of objects of interest K increases and since VN3 requires pre-computed values, VN3 suffers from the computational overhead of pre-calculating NVPs. Consequently, the performance of VN3 degenerates considerably for high densities of objects.

A solution for CNN queries: Another algorithm was proposed in [6] which provides a solution for CNN queries in road networks. Their solution is based on finding locations on a path where a NN query must be performed. CNN is based on research region principle. It uses three heuristics to generate the computation points and the region to search for the shortest path. The algorithm is brie y described and summarized in two steps: Locating a calculation point on the road and looking for the nearest point of the calculation point based on the shortest path in the road network object. The principle of this technique is to limit the search to a determined region using heuristics cited earlier. The result is in the form <point, interval, chemin> where "point" means the nearest object, "interval" is a segment of the road and "chemin" represents the shortest path from the interval to the point. The main shortcoming of this approach is that it only addresses the problem when the first nearest neighbor is requested (i.e., continuous 1-NN) and does not address the problem for continuous k NN queries.

UBA solution: The authors in [7] presented a solution called UBA for CNN queries. UBA is among the first algorithm that successfully solved the problem of

searching the KNNs in spatial network databases (SNDB). Generally, in previous works the KNNs queries and their resolution were actually based on Euclidean spaces (where the path between two objects is a straight line which connects them and which is a simple function of their spatial attributes). However, with the arrival of SNDB, objects move on predefined paths (a road for example) which are specified by a network underlying. This means that shortest network partition distance between objects depends on the network connectivity rather than the objects locations. Their algorithm is based on some properties of the algorithm IE (Intersection Examination) [7]. UBA also takes advantage of the algorithm VN3 (described above) to obtain (K + 1) NNs (k + 1) at a fixed location.

Following this purpose, UBA recovered (k + 1) of a point query q to calculate the minimum distance in which the point query can move without requiring a new issue of a NN query. UBA overcomes by far the performance of IE. However, UBA requires a large number of NN queries to determine the split points. In fact, the split point is a point in the path where KNNs of a mobile point changes. Accordingly, the execution time increases with the number and the density of objects.

IRNN solution: In [8] authors proposed four different solution methods to determine the in-route nearest neighbor (IRNN): Single graph-based (GBS), recursive spatial ranks query-based (RSR), spatial distance join-based (LDS) and area-based pre-computed (PCZ). This approach shows that the cost in terms of computational complexity, is much less expensive compared to other methods. However, comparison between the different methods showed that most of them are much more efficient when there are no updates on the road map.

Branch and Bound: Authors in [9] presented an algorithm called Branch and Bound mainly based on R-Tree solution to find K nearest neighbors objects to a point. R-Tree is the extension of Btrees in higher than one dimension. In fact, leaf nodes of the R-tree contain entries which are in the following form (RECT, oid) where oid represents the object-identifier and used as a pointer to a data object whereas RECT represents an n-dimensional Minimal Bounding Rectangle (MBR) bounding the corresponding object. As an example, in a 2-dimensional space, an entry RECT will follow this form: (xlow, xhigh, ylow, yhigh) which represents the coordinates of the lower-left and upper-right corner of the rectangle.

The Branch and Bound is based on two main metrics for the nearest neighbor search which are as follows: (MINDIST) and (MINMAXDIST). The first one is based on the minimum distance of the object O from P. The second is based on the minimum of the maximum possible distances from P to a face of the MBR which contains O. To reduce the number of nodes visited in a large search space, R-Tree algorithm uses two heuristics: search ordering and search pruning. The first is used to order the MINDIST and MINMAXDIST. And the second one is used to prune the MBR during searching. The method depends largely on the number of examples provided in entry. In this way, if the number of examples in entry is not high enough we don't expect to a good results.

CKNNs: In [10] authors proposed an algorithm which is mainly based on the Delaunay Triangulation (DT). This algorithm, applied on road networks, is able to establish the Continuous k-Nearest Neighbors (CkNNs) while taking into account

the dynamic change of locations from which queries are issued. CkNNs is an algorithm which is mainly composed of three steps: Point query localization, first nearest neighbor (1NN) determination and expansion of the search from 2NN to kNN. This step is a generalization of the previous step and tends to identify the K nearest neighbors.

The most appropriate road: The possibility of using Voronoi cells and Largest empty circle trigonometric method in the most appropriate road search was also studied in [11]. Authors proposed an algorithm which is mainly based on the Delaunay Triangulation (DT). This algorithm applied on road networks and from the current position of the doctor is able to look for the best road to the point of interest among different obstacles. After determining the destination, this algorithm is able to provide a road crossing an area without mobile obstacles. This algorithm gives good results but it still suffering from a major drawback since it is not able to treat mobiles obstacles.

3 Motivating Scenario

Improving decision-making especially in a critical case and nd the right choice in a minimum of time still always a big challenge and one of the primary concerns of the doctor. Actually, peoples die on road because of lack in communication and tracking systems and mainly because of problems and obstacles which can occur at any time. We aim by the system's architecture we present below to present our approach aiming to solve the problem of sudden death on the road. Indeed, by the conducted research, we present an emergency system able to look for the Safe Region.

We mean by this later the area without both of mobile and fixed impediments on the road. The idea behind this is to provide a road without any problem to act in a hasty manner to rescue the suffering patient. Our system is mainly based on a sub-system which is in turn based on SRAlgorithm. We present here our system architecture and in a later stage we will present the SRAlgorithm.

4 Finding the Safe Region: Proposed System's Architecture

If the doctor is available he starts looking the point of interest going by the road without any kind of obstacle neither fixed nor mobile. In fact, as we have already quoted it is highly probable that the doctor found a road without obstacle allowing him to reach quickly the point of interest, falls in a zone filled with problems and with mobile hindrances which block his path, what makes him turning in a closed circuit. The routing system presented in this paper aims firstly to look for a Safe Region devoid of any kind of obstacles to search within it the most appropriate road. The proposed system has a structure similar to the one depicted in Fig. 1.

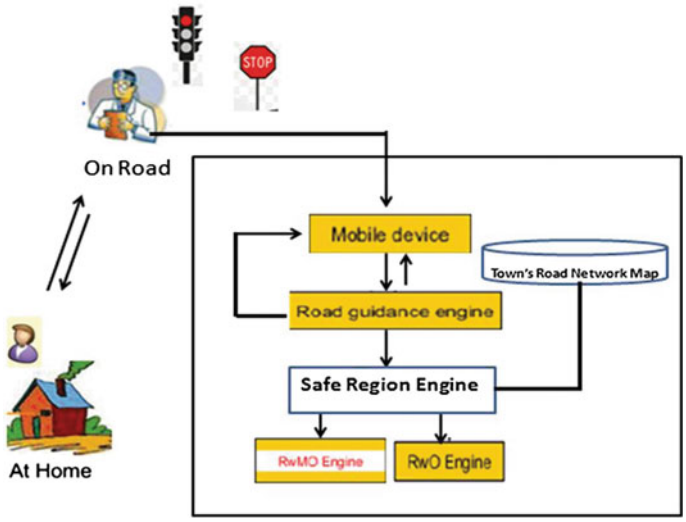


Fig. 1 Safe Region proposed system architechture [12]

The scenario is as follows: the doctor is in a given location from which he queries our system via his Mobile equipment (a PDA tablet for example) following a call from his patient. In this case the doctor is asked to look for a Safe Region to cross it and attain as quickly as possible his health care institution destination. The second component is the Road Guidance Engine responsible for looking for the most appropriate road with the help of the third component Safe Region Engine responsible for looking for the area without any kind of obstacles, taking as input a town's road network map. The research for the Safe Region is conducted by two main sub-components which are: RwMO Engine and RwO Engine.

The second one has been achieved in our previous work, which is able to look for an area without fixed obstacles. Regarding the RwMO Engine, it is able to look for an area without mobile objects/or obstacles circulating on the same area of the doctor trying to reach a patient in need. This component is based on the Safe Region algorithm that we will detail it in a later stage. Once the Safe Region has been determined, it returns a response to the Road Guidance Engine and this later takes this result as input and triggers the research of the road crossing the Safe Region to display it to the doctor on his mobile equipment.

Safe Region determination:
We assume mobile objects (obstacles) move on the map around the circle and they have not yet crossed. We assume objects move in a piecewise linear manner. In this scenario, an object moves along a straight line till it changes the direction. We use a vector $x = (x_1, x_2, x_3, \ldots. x_n)$ to denote the location of a moving obstacle. In the moving object framework, the Safe Region problem is defined as: Given a set of moving obstacles MO circulating on a map containing some fixed obstacles FO at some time instant, find the Safe Region SR to choose the right road from the current

position P at the instant t. On the map a set of fixed obstacles was already present, and we assume that this area is already devoid of these moving objects after drawing the Largest Empty Circle.

The trace of the moving objects is a set of curves, and since it is quite difficult and complicated to treat curves, we segment this set of curve on a set of segments to facilitate their processing. Recall that on the area without moving objects there are many candidate circles, so we will start searching from the largest empty circle LEC then we examine all the candidate circles (Fig. 2).

In fact a mobile object circulates on the map by drawing a trajectory that has been transformed to a set of segments. A region is so called Safe as long as a moving object does not cross the circle. In other words, moving obstacles circulate in all directions, when the object enters the circle, an event is marked at this moment and we can say that the area is not Safe, so we search the latter in another circle. Upon to the above assumptions: The set of events is defined as E. The time instances when the mobile object enters the region or the circle are called Critical Event, whose set is denoted as CE, thereby E CE. According to the Table 1, regarding the LEC, $E = t_1, t_{1+5}$, as for $CE = t_{1+10}, t_{1+15}, t_{1+20}$. Actually, the LEC and all candidate circles are drawn on fixed objects. Undoubtedly, moving obstacles are also present on the map and they change continuously their directions and can cross the road by it the user decide to go making him blocked during a certain interval of time. That's why our idea is to determine a Safe Region without neither fixed nor mobile obstacles. The idea is that upon the principle of the LEC trigonometric method, all drawn circles contains no fixed object inside, or it will not be the case if the moving objects appear on the map.

As mentioned, when the moving object enters in the circle, the status of this area changes. In fact, this event will invoke another event that is the change of the status of the Safe Region as indicated in the following table. So to check the status of the area and ensure that it is the right road, we just repeat the tracing of the largest

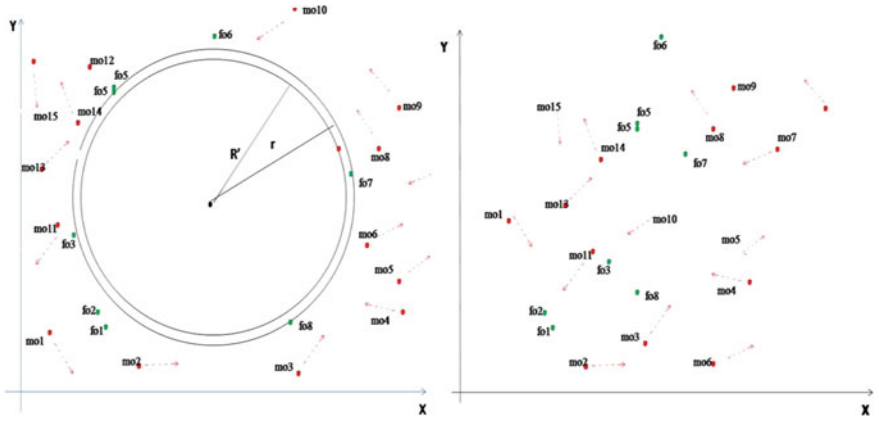


Fig. 2 Mobile obstacles circulating on original space and the appearance of new radius due crossing the LEC [12]

Table 1 The safe region/evenement figure

…	LEC	CC1	CC2	CC3
t_1	Safe	Safe	Not Safe	Safe
t1+5	Safe	Safe	Not Safe	Safe
t1+10	Safe	Not Safe	Not Safe	Safe
t1+15	Not Safe	Not Safe	Not Safe	Safe
t1+20	Not Safe	Not Safe	Not Safe	Safe

empty circle and we check if the radius of the circle has changed or not. Indeed, the advantage of this trigonometric technique is that it guarantee that no point on the map regardless of its nature, mobile or landline can be present in the circle. In this way, a moving object can be detected following a new tracing of circles. It is therefore sufficient to seek the new radius $R^{'}$ following the entry of the first moving object. If the radius $R^{'} < r$, we detect an event and the status of this region changes.

At first, all regions have the status Safe, since they contain no fixed obstacle. However, the moving object changes continually its position, but we just need it's first position to say that this area is not Safe. The time the status of the region changes, we move on to look for another Safe Region on other circles. From the figure and the following table we can check if this region is Safe or not. In reality, we do not need to know whether the moving obstacle that crosses the circle advance or it keeps its position or even if it fell, because once it has entered into a circle that was already ranked among the Safe Regions through which the user has decided to take his way, this circle in general and this region in particular has become Not Safe. This is what we summarize by the following algorithm. The first check is done on the largest circle since it already contains the most suitable route and it is considered as the largest region and the safest area. If this is not the case, the n checks are made on candidate circles and an R-tree is utilized to index these candidate circles. Obviously, when new moving obstacles appear, the Safe Region changes and we need to update this R-tree.

The idea is to find an area without any problem or a point which pose a passage impediment. As shown in the figure and explained by the algorithm, the Voronoi diagram VD (P), the Delauney triangulation DT (P) and the LEC are already drawn on the map, same for candidate circles we proceed to check whether a moving object intersect a circle in a point. This will serve to verify if the moving point has crossed the area or not, even to verify if the point which has come closer to the circle changes the direction away from the safe area. So we only need the first position while entering the circle, at this time we re-draw the LEC with the same set of fixed points and mobile objects which intersect the LEC in a set of abscissa. Then we look for the new radius R' and we compare it with the former radius r. Recall that the first characteristic of this trigonometric method is that it always give a circle which does not contain any point. So, if there is a point that crosses the circle, automatically the radius will decrease. Note that the point is in movement and there is no need to keep all its positions, we only keep the first position after the point of intersection (tangent) with the circle because the fact that this point has entered, this area it becomes Not Safe. We are therefore faced with two scenarios:

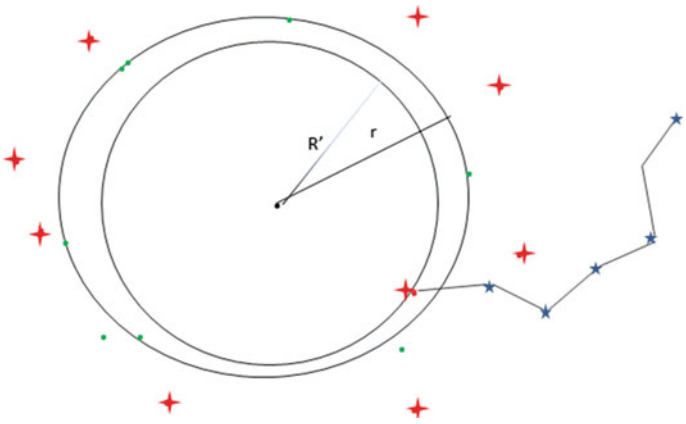


Fig. 3 Moving object crossing LEC and the appearance of new radius [12]

If the radius R = r: the area is Safe.

If the radius R$^{'}$ < r: the area is Not Safe and we have to find a safe area in one of the candidate circles and the same process is applied (Fig. 3).

Algorithm 1 CCLEC algorithm

Input: G: Road networks map, P: Set of fixed points in the map
Output: CCTree: The LEC's candidate circles, LEC : The largest empty circle

```
Compute the Voronoi diagram VD (P) and the Convex hull CH(P)
for each Voronoi vertex v do
    if v is inside CH then
        Compute radius of the circle centered on v and update max
    end if
end for
for each Voronoi edge e do
    if v is inside CH then
        Compute p = intersection between e and H and Compute radius of the circle
        centered on p and update CCTree
    end if
end for
return LEC= CCTree.max(), return CCTree;
```

Algorithm 2 Safe Region algorithm

Input: G:Road networks map, K:Set of mobile points entering the map, CCTree, LEC
Output: Safe Region(Without mobile/fix obstacles)

```
if LEC.R'=r then
    SRegion=LEC, return "Safe Region"
else if CRTree.CC.R'< r then
    return "Not Safe"
else
    SRegion=CCTree.CC, return "Safe Region"
end if
return SRegion
```

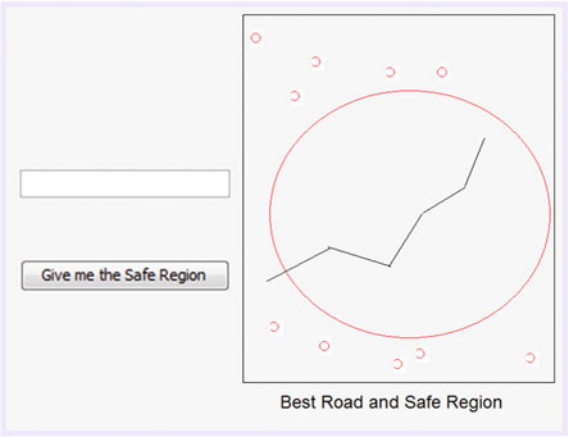


Fig. 4 Safe region pervasive system

Computing is repeated every 15 min; all states are stored in an RTree. According to all the stored states we can determine if the region is Safe or not. We begin searching the Safe Region in the LEC then in all candidate circles. In the worst case, if both of LEC and all candidate circles are not safe, the system delete these states stored in the RTree, update this later and trigger a new computing operation. Figure 4 is an overview of our pervasive system to run the scenario described above. Suppose the doctor receives a call from a patient, he is therefore called to look for the best road crossing a safe region.

First of all, as indicated in the capture screen the doctor begin by specifying his current position, then he clicks on the button "Give me the Safe Region". At this time, the road crossing the safe region is displayed to the doctor on his mobile equipment as the figure shows.

5 Concluding Remarks

The system described in this paper has been designed to help the doctor move faster through the street network of an area full of obstacles and impediments. Its role is to provide a Safe Area to guide the doctor in the shortest time possible to the site of the emergency call to save his patient's life. Our principle was to prevent the doctor to cross a region full of problems predominantly moving ones which are constantly changing and may at any time impede the doctor's passage to its destination. The proposed system implements a routing algorithm aiming to determine a Safe Region. Actually, the research of a Safe Road is conducted on the area containing mobiles points which still risk crossing the region that does not contain fixed problems. Once it has been found, it will be displayed to the doctor who already specified his destination.

References

1. Shahabi, C., Kolahdouzan, M.R., Sharifzadeh, M.: A road network embedding technique for k-nearest neighbor search in moving object databases. GeoInformatica, 94–100 (2002)
2. Shekhar, S.: Processing in-route nearest neighbor queries: a comparison of alternative approaches. In: GIS'03: Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems, pp. 9–16 (2003)
3. Tao, Y., Papadias, D.: Time-parameterized queries in spatio-temporal databases. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of data, SIGMOD'02, New York, NY, USA, pp. 334–345 (2002). http://doi.acm.org/10.1145/564691.564730
4. Papadias, D., Zhang, J., Mamoulis, N., Tao, Y.: Query processing in spatial net-work databases. In: Proceedings of the 29th International Conference on Very Large Data Bases, VLDB'03, vol. 29, pp. 802–813. VLDB Endowment (2003). http://dl.acm.org/citation.cfm?id=1315451.1315520
5. Kolahdouzan, M., Shahabi, C.: Voronoi-based k nearest neighbor search for spatial network databases. In: VLDB, pp. 840–851 (2004)
6. Feng, J., Watanabe, T.: Search of continuous nearest target objects along route on large hierarchical road network. In: Proceedings of the 6th, IASTED International Conference on Control and Application, pp. 144–149. Acta Press, Calgary (2004). http://www.jaist.ac.jp/DEWS2003/download/dews2003/2-B/2-B-04.pdf
7. Kolahdouzan, M., Shahabi, C.: Voronoi-based k nearest neighbor search for spatial network databases. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB'04, vol. 30, pp. 840–851. VLDB Endowment (2004). http://dl.acm.org/citation.cfm?id=1316689.1316762
8. Papadias, D., Zhang, J., Mamoulis, N., Tao, Y.: Query processing in spatial net-work databases. In: Proceedings of the 29th international conference on Very Large Data Bases, VLDB'03, vol. 29, pp. 802–813. VLDB Endowment (2003). http://dl.acm.org/citation.cfm?id=1315451.1315520
9. Roussopoulos, N., Kelley, S., Vincent, F.: Nearest neighbor queries. SIGMOD Rec. **24**(2), 71–79 (1995). http://doi.acm.org/10.1145/568271.223794
10. Khayat, M., Akaichi, J.: Incremental approach for continuous k-nearest neighbours queries on road. Int. J. Intell. Inf. Database Syst. **27**, 204–221 (2008)
11. Faiez, H., Akaichi, J.: Pervasive system for searching the appropriate road: a mobile physician on road network case study. In: International Work-Conference on Bioinformatics and Biomedical Engineering, p. 163 (2014)
12. de Berg, M., Cheong, O., van Kreveld, M., Overmars, M.: Computational Geometry: Algorithms and Applications, 3rd edn. Springer-Verlag TELOS, Santa Clara, CA, USA (2008)

Part III
Tools and Applications

An Effective Collaborative Filtering Based Method for Movie Recommendation

Rafał Palak and Ngoc Thanh Nguyen

Abstract Collaborative filtering approach is one of the most widely used in recommendation processes. The big problem of this approach is its complexity and scalability. This paper presents an effective method for movie recommendation based on collaborative filtering. We show that the computational complexity of our method is lower than one known from the literature, worked out by Lekakos and Caravelas (Multimedia Tools Appl 36(1–2):55–70 (2006), [10]).

Keywords Movie recommendation · Collaborative filtering · Effective recommendation

1 Introduction

Every day the world creates countless hours of movies. Every day a lot of people look for interesting movies to watch. Because of the growing number of movies this task becomes more and more difficult. People would like to see as many as possible interesting movies and as little as possible worthless movies. Besides, the same movie can be valuable for someone but worthless for another one. Therefore, personalized content is so important in the present world. The fact that we observe a significant increase of personalization need in the past few years justifies the development of recommendation methods. Therefore, in recent years the popularity of recommender systems has been growing constantly. A recommended system has been defined as "software tools and techniques providing suggestions for items to be of use to a user." [15]. Recommender systems are used for various items such as movie, music, articles etc. [15]. This article concerns the problem of movie rec-

R. Palak (✉) · N.T. Nguyen (✉)
Department of Information Systems, Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wrocław, Poland
e-mail: raf_pal@live.co.uk

N.T. Nguyen
e-mail: ngoc-thanh.nguyen@pwr.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_13

ommendation: For a user u it is needed to recommend a list of movies S which is a subset of all the movies from set M.

In [13, 16] the authors have distinguished two types of collaborative filtering approach:

- Movie rating based on prediction values—recommendation for a user based on predicted movie rating. Movie rating predictions are generated based on ratings of users similar to the active user that prediction refers to. A predicted value expresses expected level of interest in film by active user. Predicted rating is in the same scale as scale of rated movies.
- Movie rating based on prediction class—recommendation is based on the assumption that only a certain part of watched movies is positively received by user, therefore collection of movies are divided into two subsets: *Liked_movies* and *Disliked_movies*. Firstly classification of users is done due to some features e.g. demographic characteristics. Next to the active user are recommended those movies which belong to subsets *Liked_movies* of users from the same group as the active user.

The first approach is more popular, it is used in works [5, 6, 9, 10], but its disadvantage is that it usually requires higher computational complexity than the second. The first approach can be characterised by the following three steps:

1. Calculating distances between the active user and other users in the database;
2. Calculating predictions for movies ratings;
3. Choosing list of movies with the highest rating.

Movies rating prediction is a step that requires high computational complexity mainly because of the first step. Besides, movie rating prediction usually does not take into account any external factors influence rating such as mood of user or when (or with whom) a movie was watched. It is impossible to take into account all external factors and therefore rating prediction is always associated with an error. This fact makes an attempt to predict ratings difficult, and along with the large computational complexity it would be best to omit this step in determining the best solution. In [10] the authors first calculate distance between the active user and other users in the database. In this step distances are calculated for all users in the database, what causes significantly extend time of calculations for a large database. Users with a positive correlation are defined as similar users. Basing on ratings of these users, rating predictions are calculated for the active user. Some methods can reduce the computational complexity of the first approach, e.g. in [9] the authors reduce computational complexity of first step by use user clustering. Clustering of users reduces amount of potentially similar users. This makes calculations more effective. The first approach returns a list of movies with assigned predicted ratings, this allows to specify which movie is more appropriate for user than other, but calculation of ratings predictions makes the first approach slower than the second one.

The second approach is rarely used in the literature. In [2] the authors define a threshold for setting 1/4 of user ratings. Users classify these ratings as set *Liked_movies* of their preferred movies, and the remaining 3/4 of ratings is included in set *Disliked_movies*. Users are then clustered based on attributes. Users belonging to the same cluster recommend each other movies from their subsets *Liked_movies* (that is movies that are rated above the threshold). This approach does not require ratings prediction, owing to this it increases the effectiveness of algorithms based on this approach.

The most commonly used approach to resolve movie recommendation problem is to combine content based and collaborative filtering approaches. Authors of [10] represent a typical approach for this idea. For users with a small amount of ratings the content based approach is used which is independent of ratings amount. For users with greater amount of ratings collaborative filtering approach is used. Collaborative filtering approach gives better results for users with higher amount of ratings, so it is important that the recommendation method is well scalable. These two approaches complement each other very well. Collaborative filtering approach is associated with higher overhead time, therefore often for calculating recommendation the content based approach is used, especially content based ratings can be calculated offline.

This paper focuses on increasing effectiveness of collaborative filtering approach by reducing number of users who may be similar to the active user and predicting class of movie rating rather than predicting the exact value. The effect of reduced amount of users was achieved owing to the fact that users watch movies from small group of genres, therefore it is very likely that users interested in the same genres would be similar to each other. Based on this observation good scalability and effectivity can be obtained.

Finding users interested in the same genres reduces amount of possible similar users, owing to this our method does not require large computational complexity to calculate distance between users. This fact allows for quick recommendation based on collaborative filtering approach. Proposed approach does not require prediction of exact value of movie rating. This has been achieved owing to the assumption that movies rated above median of user ratings should belong to set *Liked_movies*. This assumption simplifies the process of recommendation. Reason of this fact is that this approach does not require to predict the exact value of movie rating, but only to predict the class of movie rating. By avoiding calculating exact predicted value of movie rating the computational complexity is not large the recommendation algorithm is more effective. Our approach achieves good results for minimum 20 ratings, below 20 ratings the content based approach achieves better results. This proposed method can significantly reduce the group of candidates for similar users (even 700 times).

The remaining part of this paper is organized as follows: the next section described different approaches to recommend movies. In Sect. 3 an overview of recommendation methods is presented. Section 4 includes the experiment plan and results. In Sect. 5 we present the comparison of our results with those generated by

Lekakos's and Caravelas's algorithm described in [10]. Some conclusions are included in the last section.

2 Related Works

Movie recommendation problem is a well-known and frequently encountered problem in the literature. Most often used approach is to solve this problem by combining collaborative filtering and content base approaches. In this section we focus on the collaborative filtering approach. Classical approach is described in [10], users similarity between users X and Y is calculated by the following function:

$$d(X,Y) = \frac{n\sum_i^n X_i Y_i - \sum_i^n X_i \sum_i^n Y_i}{\sqrt{n\sum_i^n X_i^2 - \left(\sum_i^n X_i\right)^2}\sqrt{n\sum_i^n Y_i^2 - \left(\sum_i^n Y_i\right)^2}} \tag{1}$$

Where X_i,Y_i are user ratings, n is number of commonly rated movies by users X and Y. Basing on users with positive correlation, predicted ratings are calculated by formula:

$$K_i = \bar{K} + \frac{\sum_{J\ \epsilon\ Neightboors}(J_i - \bar{J})r_{KJ}}{\sum_J |r_{KJ}|} \tag{2}$$

Where K_i is prediction for movie i, $\bar{K}$ is the average mean of active user ratings, $\bar{J}$ is average of similar users ratings, r_{KJ} is distance between users. For active users with small number of ratings content-based filtering is often used. To do this authors use the cosine function for measuring the similarity between vectors. Vectors contain information about all movie features like genre, actors, directors, tags etc. Collaborative filtering approach used in these methods despite their simplicity require high computational complexity. One of the reasons of this is the step for calculating distances between users. The cost of this step is dependent on the number of users in the database. The reason for greater cost is the fact that these distances are calculated between the active user and each user in the database. The cost for large databases is very high. In the next step predicted ratings for user unwatched movies are calculated based on similar users ratings. This step also requires complex calculations. The complexity of this step increases along with the increase of the number of movies in the database and the number of similar users. Authors usually apply some improvements for the method described above. In [3] the authors improve the performance by analyzing contextual factors. In [6, 12] the authors use both approaches (content base and collaborative filtering) at the same time. Often some appropriate weights are used for representing the importance of ratings (with the increase of ratings number the importance of context based approach decreases).

In [1, 4, 5] the authors present recommendation methods based on artificial immune system. Paper [8] analyses the complexity of artificial immune systems and its strong capacity to information processing characteristics, such as pattern recognition, feature selection, memory recall and learning. A good example of a system based on artificial immune network is presented in [5], where the authors define antigens as:

$$AG = \{Ag_1, Ag_2, Ag_3, \ldots, Ag_v\} \tag{3}$$

where v is number of users. Antigen contains two types of information: user ratings and information about user such as age, gender, career etc. Antibodies are generated to react antigens. Antibodies are generated in so called immunological networks. The first step in recommendation process is to generate immune networks and to calculate the affinity between antigen and immune network. If the distance is lower than the threshold antigens expand immune system. The first network is a randomly generated, other networks are generated for antigens with worst affinity to existing immune networks. This step is performed until each antigen will belong to at least one network. The last step is the rating prediction, the authors propose to calculate the similarity between users belonging to the same immune networks and basing on them they calculate the rating prediction. Users with similarity greater than the threshold are defined as similar users. Base on that set of users the prediction is performed. For this purpose the following formula is used:

$$P_{u_v,r} = \overline{u_v} + \frac{\sum_{i=1}^{U} GSim_{revised}(u_{v,g}, u_{i,g}) \times Sim_{revised}(u_v, u_i) \times (u_{i,r} - \overline{u_i})}{\sum_{i=1}^{U} \left| GSim_{revised}(u_{v,g}, u_{i,g}) \times Sim_{revised}(u_v, u_i) \right|} \tag{4}$$

where $\overline{u_v}$ and $\overline{u_i}$ are ratings of users u_v and u_i, $GSim_{revised}$ and $Sim_{revised}$ are similarity groups. This approach requires repeated calculation of the distance between immune systems and antigens, therefore this step is very expensive. In the next step the system calculates rating prediction. This step requires complex calculations but it is less expensive than the approach in [10] because it is based on a smaller set of similar users. Each step requires high computational power or large amount of time. The first step can be calculated offline, this may slightly increase efficiency of the algorithm.

Another approach to solve the recommendation problem is to find attributes of the movie (genre, director, actors, etc.) that have the greatest impact on rating. In [14] the authors calculate the weight of each attribute based on the frequency of its occurrence in rated movies. In [9] and [7] the authors assume that the most important attribute is the genre. Authors of [9] cluster users basing on the genre preferences and demographic data (age, gender, occupation). For users belonging to the same group of the active user distances are calculated between them and the active user. The next step includes rating prediction. Authors used Eq. (2) to predict movie ratings. Top N movies with the highest predicted ratings are recommended. This method requires a large number of operations for clustering users. The step of

predicting movie rating is time-consuming because it calculates rating for each unwatched movie.

All presented above, recommendation methods require large computational complexity. This makes it impossible to provide recommendation in an effective way, especially in real-time systems.

3 Our Proposed Recommendation Method

We assume that each user can independently rate movies in scale 1-10. For some users rating 8 could mean good movie while for others it is a poor movie. Therefore, each user needs an individual approach for recommendation. To achieve the individual approach for each user it is necessary to define a threshold value, movies rated above this threshold value will be classified as good movies. In this paper the threshold value is assumed to be equal to the median of all user ratings. Movies rated above the threshold will belong to the set *Liked_movies*. Later on in this paper the expressions "*movies rated above the median*" and the term "*movies liked by the user*" will be used interchangeably. Stating that movies rated above median give user the greatest satisfaction it is not necessary to predict exact value of movie rating, therefore in this paper we use movie rating based on prediction class approach described in the Introduction. In this paper we assume that recommended movies must belong to left-open interval (*median value of user ratings*, *max rating*].

The majority of movies watched by users belong to a few genres. Each user has a few genres that likes to watch most. Based on this assumption we can have the following observation: Users who watch the same genres movies more likely will be similar to each other. Using these assumptions we formulate the steps of the recommendation algorithm as follows:

1. Group rated movies by genre;
2. Calculate the weight for each genre and select N best genres;
3. Find users for whom the set of liked genres is equal in a certain extent to the set of liked genres of active user;
4. Calculate distances between users in groups;
5. Select similar users whose distance is less or equal to the threshold value;
6. Create the recommendation list of movies sorted by weighted frequency of occurrence in similar users liked movies set.

To calculate the weight of genre the following function is used:

$$w = \frac{2s + g}{3a} \tag{5}$$

where g is the number of movies that belong to particular genre rated by user, s is the number of movies belonging to a particular genre rated above the median, and a is the number of all movies rated by the active user. This function returns a value in interval $< 0,1 >$. The higher value, the more popular genre. Note that above function rewards particularly genres of liked movies because it is fairly possible that highly rated movie belongs to disliked genre. The second component of this equation is the number of movies from a particular genre. This component rewards the most common genres in user ratings because large number of movies in this genre suggests that the user likes a particular genre. This component is very important for users with small number of ratings because it is very likely that user might not rate good movies of this genre yet. Based on calculated values of the function we are able to determine what genres user like most. It is a key step for this algorithm. Based on a set of favorite genres databases are searched for similar users liked the same genres (set of best N genres overlaps to the desired extent to set of best N genres of active user). For each similar user the distance between active user is calculated by formula:

$$d(u_1, u_2) = \begin{cases} 1 & for\ \left(\left(A' \cap B\right) \cup \left(B' \cap A\right)\right) = \varnothing \\ 1 - \dfrac{\dfrac{\left|A' \cap B'\right|}{\left|\left(A' \cap B\right) \cup \left(B' \cap A\right)\right|} + \dfrac{\left|A'' \cap B''\right|}{\left|A'' \cup B''\right|}}{2} & for\ \left(\left(A' \cap B\right) \cup \left(B' \cap A\right)\right) \neq \varnothing \end{cases} \quad (6)$$

where A and B are set of movies watched by users u_1 and u_2, respectively, 'A' and 'B' are sets of movies liked by users (movies rated above median or rated by maximal rating) u_1 and u_2, 'A' and 'B' are sets of N best genres for users u_1 and u_2. This function is a semi-metric. The distance is based on two aspects: the compatibility of favorite genres and the compatibility of ratings. The next step filters users for experimentally determined threshold, users with distance equal or below the threshold are assumed to be similar to each other. The last step performs recommendation of movies liked by similar users to the active user. Movies appearing greatest number of times are most important for the active user but this aspect has been omitted from the study because limitations of the research method.

The algorithm described above may be used only for users with at least 20 ratings. For users with less than 20 ratings a content-based approach algorithm can be used successfully, for example cosine similarity measure. Each movie is represented by a vector describing all movie features like genre, actors, writers, directors, then movies with lower distance to movies rated by the active user are recommended. We used this method to determine the minimal number of ratings for which the recommendation based on the proposed method is better than the content based approach.

4 Experiments

We have used MovieLens data set which contains 30,000 movies, 230,000 users and 21,000,000 ratings in scale 1–10. This data set is available at [18]. Data were complemented with information about cast and crew from [19]. Users were divided into two groups: the first contains 60,000 users whose movies are recommended, and the second contains 150,000 users for which similar users are found.

For each active user for whom the recommendation was performed, 10 % of ratings (at least 1) was hidden. These movies do not participate in the recommendation process. Users are treated as if he/she never rated them. It is necessary to check performance of the algorithm. Recommendation was made on the basis of the rest of user ratings, next the results of recommendations were compared with a hidden part of ratings.

The most commonly used measure of the recommendations requires rate prediction. Therefore, to evaluate the results two measures, precision and recall have been applied. Precision is defined as:

$$precision = \frac{recommendedMoviesAboveMiedian}{allHiddenMoviesAboveMedian} \tag{7}$$

where *recommendedMoviesAboveMiedian* is the number of all recommended movies rated by users above the median, and *allHiddenMoviesAboveMedian* – the number of all hidden movies rated by users above median. Recall is defined as:

$$recall = \frac{recommededMoviesLessOrEqualMedian}{allRecommendedMovies} \tag{8}$$

where *recommededMoviesLessOrEqualMedian* is the number of all recommended movies with ratings equal or less than the median, if the median is not an integer then ratings median ±0,5 are counted as equal median, *allRecommendedMovies* is the number of all recommended movies with known ratings. The results are presented in Table 1.

Table 1 Results of recommendation

Minimal number of ratings	Recall for 0.0 (%)	Precision for 0.0 (%)	Recall for 0.01 (%)	Precision for 0.01 (%)	Recall for 0.1 (%)	Precision for 0.1 (%)
20	0	49.02	0.42	49.12	68.68	68.68
30	0	39.22	0.54	39.35	68.24	68.24
40	0	32.79	0.66	32.95	67.79	67.79
50	0	27.28	0.82	27.48	67.29	67.29
60	0	22.44	1.01	22.69	66.78	66.78
70	0	18.33	1.21	18.6	66.24	66.24
80	0	14.94	1.42	15.25	65.79	65.79
90	0	12.23	1.58	12.58	65.37	65.37

In Table 1 the first column represents the minimal number of ratings for users to whom the recommended movies are, they took part in the recommendation as similar users. Other columns represent the results of both measures for different thresholds used in the fifth step of the algorithm. During the study the most important measure was the recall because users quickly lose confidence in system which deliveries poor recommendations. We can note that the best results were achieved of recall for threshold equal 0.0. For a distance equal 0.0, compatibility of favorite genres must be equal to 100 %. Because of that the third step of the algorithm could be reduced to find users with the same set of liked genres. Note that studies are conducted for a minimum of 20 ratings because for users with less number of ratings than 20 this method provides unsatisfactory results. Content based approach is much better in this situation.

5 Comparison Results

To examine the efficiency of the solution, we have used the method described in [10] for comparison with our approach. We have decided to choose the method worked by Lekakos and Caravelas [10] because of its popularity and its practical aspect. In both solutions the same content based approach is used, therefore, comparison focuses on comparing both collaborative filtering components. Both approaches are significantly different in the ways to realize recommendation. In [10] the authors are trying to predict movie rate, our proposed method focuses on recommending movies above median without predicting exact values of movie rating. Comparison of both algorithms requires some modification for approach described in [10]. Therefore, result returned by [10] are filtered first. Movies which predicted rating above the median are removed. The list of filtered movies can be treated as a set of returned by our algorithm. Note that both algorithms predict movies rated above median. The comparison results are presented in Table 2.

In Table 2 the first column represents the minimal number of ratings for users for whom the recommended movies are, that is users who took part in the recommendation as similar. The remaining two columns represent the results of both measures.

Table 2 The results of algorithm described in [10]

Minimal number of ratings	Recall (%)	Precision (%)
20	78.17	34.23
30	78.03	34.2
40	77.74	34.54
50	77.56	34.34
60	77.43	34.8
70	77.39	34.72
80	77.21	35.06
90	77.09	35.65

We can note that this algorithm achieves much worse results for recall than proposed in this paper algorithm. For threshold equal 0.0 and users with at least 20 ratings algorithm proposed in [10] reaches worse result than presented algorithm but presented algorithm for threshold equal 0.0 precision value decreases with increasing amounts of ratings, in other hand for algorithm presented in [10] precision value is constants and for minimum 40 ratings algorithm achieve better result than results for threshold equal 0.0. For threshold value of 0.1 our proposed algorithm gives much better results for recall and slightly better results for precision than the algorithm proposed in [10]. Thus our algorithm is better not only regarding the computational complexity but also gives better results. In other words, our algorithm is more effective.

6 Conclusions

In this paper an effective method for movie recommendation is presented. Despite the lower computational complexity of method, the proposed algorithm gives better results than one known in the literature. The algorithm achieves good results for both measures: recall and precision. Despite the good results the algorithm still requires further research, special attention should be paid to the dynamics of user preferences in time. Ontology-based approaches can be useful for this purpose [17, 11].

References

1. Acilar, A., Arslan, A.: A collaborative filtering method based on artificial immune network. Expert Syst. Appl. **36**(4), 8324–8332 (2009)
2. Basu, C., Hirsh, H., Cohen, W.: Recommendation as classification: using social and content-based information in recommendation. In: Proceedings of the 15th National Conference on Artificial Intelligence, pp. 714–720. Madison, WI (1998)
3. Biancalana, C., Gasparetti, F., Micarelli, A., Miola, A., Sansonetti, G.: Context-aware movie recommendation based on signal processing and machine learning. In: Proceedings of the 2nd CAMRa '11, pp. 5–10. ACM, New York, NY, USA (2011)
4. Cayzer, S., Aickelin, U.: A Recommender system based on idiotypic artificial immune networks. Algorithms **4**, 181–198 (2005)
5. Chen, M., Teng, C., Chang, P.: Applying Artificial immune systems to collaborative filtering for movie recommendation. Adv. Eng. Inform. **29**, 830–839 (2015)
6. Chikhaoui, B., Chiazzaro, M., Wang, S.: An improved hybrid recommender system by combining predictions. In: Proceedings of IEEE Workshops of International Conference on Advanced Information Networking and Applications, pp. 644–649 (2011)
7. Choi, S., Ko, S., Han, S.: A movie recommendation algorithm based on genre correlations. Expert Syst. Appl. **39**(9), 8079–8085 (2012)
8. Dasgupta, D., Ji, Z., Gonzalez, F.: Artificial immune systems research in the last five years. In: Proceedings of the Congress on Evolutionary Computation Conference, pp. 8–12. Canberra (2003)

9. Kyung-Rog, K., NamMee, M.: Recommender system design using movie genre similarity and preferred genres in smartphone. Tools Appl. **61**(1), 87–104 (2011)
10. Lekakos, G., Caravelas, P.: A hybrid approach for movie recommendation. Multimedia Tools Appl. **36**(1–2), 55–70 (2006)
11. Maleszka, M., Mianowska, B., Nguyen, N.T.: A method for collaborative recommendation using knowledge integration tools and hierarchical structure of user profiles. Knowl.-Based Syst. **47**, 1–13 (2013)
12. Nguyen, N.T., Sobecki, J.: Using consensus methods to construct adaptive interfaces in multimodal web-based systems. J. Univ. Access the Inf. Soc. **2**(4), 342–358 (2003)
13. Pham, X.H. et al., Spear, A.V.: A new method for expert based recommendation systems. Cybern. Syst. **45**(2), 165–179 (2014)
14. Pham, X.H., Jung, J., Nguyen, N.T.: Ontology-based multilingual search in recommendation systems. Acta Polytech. Hung. **13**(2), 195–207 (2015)
15. Ricci, F., Rokach, L., Shapira, B., Kantor, P.: Recommender Systems Handbook, pp. 1–10. Springer, New York, NY, USA (2010)
16. Sarwar, B., Karypis, G., Konstan, J., Riedl, J: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International World Wide Web Conference, pp. 285–295 (2001)
17. Sriharee, G.: An ontology-based approach to auto-tagging articles. Vietnam J. Comput. Sci. **2** (2), 85–94 (2015)
18. http://grouplens.org/datasets/movielens/
19. http://themoviedb.org

A Linux Kernel Implementation of the Traffic Flow Description Option

Robert R. Chodorek and Agnieszka Chodorek

Abstract The Traffic Flow Description is an option of the IP protocol that allows end-systems to describe generated traffic flows. Such description includes instantaneous values of transmitted data in a given time. The option enables intermediate systems to assure QoS based on dynamic resource allocation. In this paper an implementation of the Traffic Flow Description option for the Linux kernel is presented. The paper includes both the description of the option, proposed by the Author as the Internet Draft working document and detailed description of the prototype implementation of the proposed option in the Linux kernel. The implementation covers both improvements introduced to the current long term stable 4.1.20 version of the Linux kernel and two helper functions that enable the option to be set up easily. Tests show that the functionality of the prototype implementation complies with the specification of the option, given in the Internet Draft. Results of performance tests show that the prototype implementation is able to work as a part of the system of QoS assurance.

Keywords QoS · Multimedia systems · Video streaming · Heterogeneous IP network · IP traffic flow description option

1 Introduction

The Traffic Flow Description (TFD) Option [6] is a newly developed option of the Internet Protocol (IP). The option is intended for signalling purposes, for the sake of dynamic QoS assurance. The traffic flow description, based on this option, includes

R.R. Chodorek (✉)
Department of Telecommunications, The AGH University of Science and Technology,
Al. Mickiewicza 30, 30-059 Krakow, Poland
e-mail: chodorek@agh.edu.pl

A. Chodorek
Department of Information Technology, Kielce University of Technology,
Al. Tysiaclecia Panstwa Polskiego 7, 25-314 Kielce, Poland
e-mail: a.chodorek@tu.kielce.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_14

the amount of data that will be transmitted by an application in a given time. Information about forthcoming traffic can be acquired from the sending buffer, from a traffic predictor or directly from a traffic source (e.g. from compression process in the case of audio or video transmission). Such information, when transmitted hop-by-hop from the source of the traffic to the receiver (or receivers) can be used by intermediate nodes to dynamically allocate of network resources. As an effect, the Quality of Service (QoS) assurance based on the TFD option is established, and the option itself works as signalling, being able to replace previous signalling protocols, such as the Resource ReSerVation Protocol (RSVP) [4], in the reservation process.

In modern networks, QoS assurance systems typically work in a static manner, which give them the possibility of using statistical knowledge about transmitted traffic. Although these reservations allow end-systems to achieve a satisfactory QoS [2, 10], it is done at the cost of link utilization. Thus, current research efforts in the area of QoS are focused, amongst other things, on the possible role of dynamic bandwidth allocation. Dynamic network resource allocation used to minimize response time in the Information-Centric Networks (ICN) was proposed in [3]. The solution for Component-Based Software Engineering (CBSE), which allows adaptive end-to-end resource reservation of the network resources for distributed real-time systems to allow efficient resource utilizations, is shown in [9]. A new architecture, designed for Software Defined Networks (SDN), which dynamically adapts to application requirements and, as a result, more efficient use network resources, was proposed in [5]. Results described in the paper [7] show that dynamic reservations based on the TFD option successfully deal with the trade-off between reservations accuracy and link utilization.

The aim of this paper is to present a Linux kernel implementation of the TFD option which can be used for dynamic QoS assurance. The paper is organized as follows. Section 2 describes the TFD IP option. Section 3 discusses details of our implementation of the TFD option as a modification to the Linux kernel and shows method of its usage. Section 4 shows functionality of our implementation. Section 5 concludes the paper.

2 The TFD Option of IP Protocol

The TFD option of the IP protocol is introduced by the Internet Draft working document [6]. The option can be put in an IPv4 header (Fig. 1a) or can be used as an option header (so-called extension header) of the IPv6 protocol (Fig. 1b). The format of the IPv4 and IPv6 options differ in the first byte. In the case of IPv4, the first byte identifies the option. Putting the TFD option into the IPv4 header needs the first bit set to 1, which indicates that during fragmentation of an IP packet, the TFD option must be copied into all fragments of the packet. Because the TFD option belongs to the control options, the next two bits of the first byte are set to 0. The last five bits of the first byte specify the unique option number, allocated by IANA. In the case of the IPv6 protocol, the first byte of the TFD option (*Next Header* field) will contain

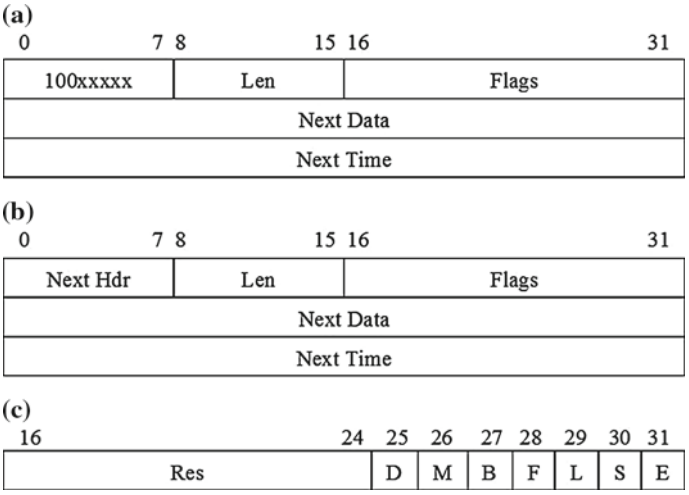


Fig. 1 Format of the TFD option [6, 7]: **a** IPv4, **b** IPv6, **c** *Flags* field

information about the type of encapsulated header that is next in line. So the *Next Header* conveys information about the option type of the next extension header or the identifier of the transport protocol.

The TFD option consists of five fields. The first field is described above. The *Len* field contains the length of the TFD option (IPv4) or the length of the TFD header (IPv6) in bytes (i.e. 12 bytes). The *Flags* field (Fig. 1c) describes properties of the *Next Data* field (*D*, *M*, *B* and *F* flags), existence of flow label (*L* flag), and general properties of transmitted traffic (*S* and *E* flags). Last but not least, *Next Data* and *Next Time* fields, are used for the description of forthcoming traffic. The *Next Data* field conveys information about amount of data (in bytes) that will be send in time given by the *Next Time*. The value of the *Next Time* is expressed in milliseconds.

3 Prototype Implementation of the TFD Option in the Linux Kernel

In the Linux operating system, network processing in the lowest four layers of the Open Systems Interconnection (OSI) model is handled by the Linux Kernel Networking stack [1, 11]. As an effect, the Linux implementation of any IP option requires improvement to the Linux kernel.

This section presents the Linux implementation of the proposed IP option. It describes improvements introduced to the Linux kernel (both to the version 4 and the version 6 of the IP protocol) to enable usage of the TFD option. It also shows how to set up the implemented option. The described implementation of the TFD option was built for the current long term stable 4.1.20 version of the Linux kernel.

3.1 Improvements in the Linux Kernel

The Linux kernel is a monolithic kernel, where all modules works in supervisor mode. Incoming and outgoing packets are represented in the Linux kernel by `sk_buff` structures. A socket buffer (SKB) stores both packet headers (from layer 2 to layer 4) and a packet payload [11].

A packet, received by a network adapter, is stored in the SKB and, next, is handled by a matching processing method—e.g. by the `ip_rcv()` method for the IPv4, and by the `ipv6_rcv()` method for the IPv6. Further processing is carried out by a number of functions, that (based, amongst other things, on analysis of addresses) will transfer the packet to higher layers (if a node address is equal to the packet destination address), or to routing procedures (if a packet must be transferred to another end system). In both situations, IP options can be processed. Functions that process IP options will be presented later.

In the Linux IP protocol implementation, the sending method depends of the version of the IP protocol and a type of the transport protocol. In the case of transport protocols that handle fragmentation by themselves, the `ip_queue_xmit()` method is used (IPv6: the `ip6_xmit()` method). In the case of transport protocols that do not handle fragmentation, the `ip_append_data()` method is used (IPv6: the `ip6_append_data()` method). Further packet processing in the Linux kernel is carried out by a number of functions that, after the IP packet has been completely built, transfer it to the network interface.

The implementation of a new IP option requires the introduction of several parameter definitions which will be used in functions that process this option. The first required parameter is the option number. In the case of IP protocol version 4, the number assigned for the IP option (here: identified by the `IPOPT_TFD`) is stored in the kernel header `ip.h` (`<uapi/linux/ip.h>`). The definition of `IPOPT_TFD` is as follows:

```
#define IPOPT_TFD (29|IPOPT_COPY|IPOPT_CONTROL)
```

The above definition consists of the option number (temporary assigned number: 29) and two flags, `IPOPT_COPY` and `IPOPT_CONTROL` [6]. The `IPOPT_COPY` flag indicates that the option must be copied to all IP fragments [11]. The `IPOPT_CONTROL` flag defines the TFD option as belonging to the control class [11].

In the case of the IP protocol version 6, the number assigned to the TFD option (parameter: `NEXTHDR_TFD`) is stored in the kernel header `ipv6.h` (`<net/ipv6.h>`) [6]. The definition of the `NEXTHDR_TFD` contains only the header number (here: temporary assigned):

```
#define NEXTHDR_TFD 29
```

The implementation in the Linux kernel also requires the definition of constants—flags, which can be set in the TFD option [6]. The definition of the flags for the option working as a part of the IPv4 protocol is stored in the `uapi/linux/ip.h` header file:

```
#define IPOPT_TFD_D   0x0040
#define IPOPT_TFD_M   0x0020
#define IPOPT_TFD_B   0x0010
#define IPOPT_TFD_F   0x0008
#define IPOPT_TFD_L   0x0004
#define IPOPT_TFD_S   0x0002
#define IPOPT_TFD_E   0x0001
```

A similar definition for IPv6 is stored in the `net/ipv6.h` file:

```
#define IPv6_TFD_D   0x0040
#define IPv6_TFD_M   0x0020
#define IPv6_TFD_B   0x0010
#define IPv6_TFD_F   0x0008
#define IPv6_TFD_L   0x0004
#define IPv6_TFD_S   0x0002
#define IPv6_TFD_E   0x0001
```

The structure with the defined format of the TFD option for IPv4 is stored in the `uapi/linux/ip.h` file:

```
struct ip_tfd_hdr {
   __u8  opttype;
   __u8  optlen;
   __be16 flags;
   __be32 next_data;
   __be32 next_time;
};
```

A similar structure was defined for IPv6 (in the file `net/ipv6.h`):

```
struct ipv6_tfd_hdr {
   __u8  nexthdr;
   __u8  hdrlen;
   __be16 flags;
   __be32 next_data;
   __be32 next_time;
};
```

In the case of received IPv4 packets, IP options are processed by the `ip_rcv_options()` function, which is called from the `ip_rcv_finish()` function. The

ip_rcv_options() function calls the ip_options_compile() method, which parses the IPv4 header of the specified SKB and builds an IP options object. The ip_options_compile() method validates the TFD option, using the following code:

```
case IPOPT_TFD:
    if (optlen < 12) {
      pp_ptr = optptr + 1;
      goto error;
   }
break;
```

In the case of received IPv6 packets, options are processed by the ipv6_find_hdr() function which is called from the ip6_input_finish() function. The ipv6_find_hdr() function, amongst other things, validates the TFD option:

```
if (nexthdr == NEXTHDR_TFD) {
   if (hdrlen <12)
      return -EBADMSG;
}
```

In routers, the TFD option is both conveyed to a QoS engine of a current router, and forwarded (as a part of IP datagram) to the next router on the path. In end-systems, upon the demand of an application, the TFD option can be delivered to this application as a typical IP option (using the getsockopt() function of the Linux kernel).

Applications which are required to send the TFD IP option must prepare the data needed by the option and process them via a standard socket Application Programming Interface (API) (details will be presented in the next sub section). In our implementation, opt_tfd is a name of a variable of the ip_tfd_hdr type. In the case of IPv4 protocol, the ip_options_build() method is called to build the TFD option. This method writes the content of the opt_tfd structure (i.e. content of the TFD option) to the IPv4 header:

```
memcpy(&(IPCB(skb)->opt), opt_tfd, sizeof(struct
                                    ip_tfd_hdr));
```

The ip_options_build() method is called for both transmission methods: with the use of the ip_queue_xmit() method or with the use of the ip_append_data() method. A similar ip_options_build() procedure was written for the IPv6 protocol and it is stored in the ip6_output.c file.

3.2 Setting the TFD Option

Applications that use the IP TFD option must prepare short-term network resource requirements for QoS assurance. Short-term resource requirements are written as two parameters, the *Next Data* and the *Next Time*, and a proper set of flags. The simplest method for evaluation of these requirements is to calculate current occupation of the application sending buffer and to estimate time of buffering.

The application must prepare all data necessary to setting up the TFD option in the temporary buffer, and then put the data into the IP module in the Linux kernel, using the standard socket API (function `setsockopt`). Setting up the TFD option from an application is described below (`rspace` is an identifier of a variable that act as a buffer):

```
setsockopt(sock_desc, SOL_IP, IP_OPTIONS, rspace, 12);
```

To simplify the usage of the TFD option, a helper function for each IP protocol version was defined. This function (`set_IPv4_TFD` for IPv4 and `set_IPv6_TFD` for IPv6) has two definitions—the first with the *Next Data* field as an integer and the second with the *Next Data* field as a floating point value (variable of the float type). The function requires socket identifier, flags, values of *Next Data* and *Next Time*. Definitions of a helper function for C++ programming language are as follows:

```
int set_IPv4_TFD(int sock_desc, uint16_t flags,
    uint32_t next_data, uint32_t next_time);

int set_IPv4_TFD(int sock_desc, uint16_t flags,
    float next_data, uint32_t next_time);
```

The implementation in C++ of the function `set_IPv4_TFD()` for *Next Data* as an integer is placed below:

```
int set_IPv4_TFD(int sock_desc, uint16_t flags,
    uint32_t next_data, uint32_t next_time) {

   char rspace[40];
   uint16_t flag_mask;
   rspace[0] = IPOPT_TFD;
   rspace[1] = 12;
   flag_mask = ~IPOPT_TFD_D;
   flags = flags & flag_mask;
   flags = htons(flags);
   next_data = htonl(next_data);
   next_time = htonl(next_time);
   memcpy(&rspace[2], &flags, sizeof(flags));
```

```
    memcpy(&rspace[4], &next_data, sizeof(next_data));
    memcpy(&rspace[8], &next_time, sizeof(next_time));
    if (setsockopt(sock_desc, SOL_IP, IP_OPTIONS,
                   rspace, 12) < 0) {
        return -1;
    }
    return 0;
};
```

In the above function, the user buffer `rspace` is filled by the data necessary for the TFD option. The first byte of the buffer contains the type of the TFD option. Next byte contains the length of the TFD option (according [6], this byte must be set to 12). Flags are set by a user, with the exception of flag *D* which must be set to 0 (for integer value in the field *Next Data*). *Flags* and *Next Data* and *Next Time* fields must be converted to network byte order and then copied to the `rspace` buffer.

Function `set_IPv4_TFD()` tries to set the TFD option using the socket API. If this operation completes successfully, the function will return 0. If the `set_IPv4_TFD()` function fails to set up the TFD option, it will return −1.

The implementation in C++ of the `set_IPv4_TFD()` function for Next Data as a floating point variable needs the declaration of the *Next Data* field as float type. In this case, flag *D* must be set to 1 (the *Next Data* field of the TFD option has a floating point value).

4 Tests and Usage of the TFD Option

A series of tests of the implementation, described in the previous Section, were performed. There were both functional tests of the implementation and performance tests of an entire QoS assurance system that applied to the TFD option. Tests of the entire system were carried out using the improved ns-2 emulator [8], and as the real-world user application have acted the modified VideoLAN Client (VLC) application (standard VLC with the TFD option implemented by the authors). Because performance tests are out of scope of this paper, the interested reader is referred to [7] for further information.

During functional tests, a user application sent data packets with the TFD option in an IP header. Transmitted IP packets were captured using a modified version of the `tcpdump`. Functional tests were conducted on a test network where end-systems and intermediate systems were built on the base on Linux systems.

Helper functions, shown in the Sect. 3.2, allow the user to simplify setting up the TFD in outgoing IP packets. Before sending the data packet, the user program calls the `set_IPv4_TFD` (`set_IPv6_TFD`) function with required arguments.

Let's assume that an application sends data in real time, and the TFD option is set up on the basis of the buffer analysis. According to these assumptions, two flags *Flags* field must be set two flags *S* and *B* (`IPOPT_TFD_S` and `IPOPT_TFD_B`)

must be set up in the Flags Field. Let's assume also that (based on the buffer analysis) the *Next Data* is set to 80000 bytes and the *Next Time* is set to 300 ms. A fragment of a program which uses the `set_IPv4_TFD()` function to setup the TFD option according to assumed data is presented below:

```
flags = IPOPT_TFD_S | IPOPT_TFD_B;
next_data = 80000;
next_time = 300;

if (set_IPv4_TFD(sk_p, flags, next_data, next_time)) {
   printf("Error...\n");
}
```

An IP packet with the TFD option set by a code fragment presented above was transmitted via Ethernet network, where it was captured by the `tcpdump`. The result of the capture is shown below (first 48 bytes of the Ethernet frame):

```
0x0000:  0800 27d8 0101 0800 27d8 a2f2 0800 4800
0x0010:  0558 2eb8 4000 0111 3497 0a00 020f 0a00
0x0020:  0211 9d0c 0012 0001 3880 0000 012c a2e1
```

The sequence of bytes "`9d0c 0012 0001 3880 0000 012c`" is the content of the TFD option. The first byte is set according to the type of conveyed option [6] and `IPOPT_COPY` flag (hexadecimal value `9d`) and the next byte denotes length of the option (12, expressed hexadecimally as `0c`). In the TFD option flag field (hexadecimal value `0012`) are set flags *S* and *B* (the streaming traffic and the *Next Data* field is set on the basis of a buffer analysis). The *Next Data* field (hexadecimal value `0001 3880`) is set to 80,000 bytes and the *Next Time* field (hexadecimal value `0000 012c`) is set to 300 ms.

5 Conclusions

This paper discussed the software implementation of the new IP option—the Traffic Flow Description option—in the current long term stable 4.1.20 version of the Linux kernel. The implementation includes both improvements made in the kernel, and helper functions to facilitate usage of the implementation. Our experimental results show that the implementation achieved expected functionality, and results reported in the paper [7] show that the usage of the implemented TFD option for signalling in the QoS assurance system can bring significant improvements to link utilization.

Our implementation of the TFD option extends (but does not alter) current network processing in the Linux kernel. Therefore, TFD-capable transmission is transparent to IP systems that do not implement the TFD option.

Acknowledgments The work was supported by the contract 11.11.230.018.

References

1. Abeni, L., Kiraly, C.: Investigating the network performance of a real-time Linux Kernel. In: 15th Real Time Linux Workshop (2013)
2. Asghar, J., Faucheur, F.Le, Hood, I.: Preserving video quality in IPTV networks. IEEE Trans. Broadcast. **55**(2), 386–395 (2009)
3. Avci, S.N., Westphal, C.: A content-based traffic engineering policy for Information-Centric Networks. In: 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp. 711–719 (2016)
4. Braden, R. (ed.), Zhang, L., Berson, S., Herzog, S., Jamin, S.: Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification. RFC 2205 (1997)
5. Bueno, I., Aznar, J.I., Escalona, E., Ferrer, J., Garcia-Espin, J.A.: An opennaas based sdn framework for dynamic qos control. In: Proceedings of the IEEE SDN for Future Networks and Services (SDN4FNS 2013), pp. 1–7. Trento (2013)
6. Chodorek, R.R.: An IP option for describing the traffic flow. Internet Draft draft-chodorek-traffic-flow-option-04.txt, IETF (2015)
7. Chodorek, R.R., Chodorek, A.: Providing QoS for high definition video transmission using IP traffic flow description option. In: Proceedings of IEEE Conference on Human System Interaction, pp. 102–107. Warsaw, Poland (2015)
8. Chodorek, R.R., Chodorek, A.: Expanding the Ns-2 emulation environment with the use of flexible mapping. In: book: Computer Networks, Communications in Computer and Information Science Series, vol. 608, pp. 22–31 (2016)
9. Khalilzad, N., Ashjaei, M., Almeida, L., Behnam, M., Nolte, T.: Towards adaptive resource reservations for component-based distributed real-time systems. ACM SIGBED Rev. **12**(3), 24–27 (2015)
10. Pana, F., Put, F.: A survey on the evolution of RSVP. IEEE Commun. Surv. Tutor. **15**(4), 1859–1887 (2013)
11. Rosen, R.: Linux Kernel Networking: Implementation and Theory. Apress (2013)

The Quality of Internet LTE Connections in Wroclaw

Jerzy Kisilewicz

Abstract Measurement results of quality LTE connections are presented in this paper. Play operator was selected as the LTE service provider. The transmission bitrates and response times were measured for home network and also while driving with a speed up to 130 km/h. The downloading and uploading bitrates for home network were observed during two weeks and there was calculated how these bitrates have changed during the day. The percentage of samples having a specific bitrates was calculated. There is shown that the request response time depends on the transmission bitrate. LTE connection stability was observed while driving.

Keywords Wireless communication · Mobile internet · LTE technology · Transmission quality

1 Introduction

Long Term Evolution (LTE) standard, which is a huge improvement in wireless data transmission, is the successor of 3G systems. Expected downstream bitrate amounts up to 300 Mbps, and upload bitrate to 50 Mbps. In LTE, the respose time could be reduced to approx. 10 ms and managed to increase the resilience and resistance to interference, as described by Dahlman [1, 2]. LTE connections maintain good parameters even while moving at a speed of over 100 km/h, although they are optimized for low speeds up to 15 km/h, as defined in the technical white paper [3]. The development of LTE is very dynamic now. Future development of LTE in the US show the 4G Americas white papers [4, 5].

The LTE services in Poland provide operators: Play, Plus, Orange, T-Mobile and Cyfrowy Polsat. It was used only the 1800 MHz frequency, but in December, 2014

J. Kisilewicz (✉)
Chair of Computer Systems and Networks, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: jerzy.kisilewicz@pwr.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_15

the operator of Play [6] launched the ability to use LTE in the Warsaw subway, and for the first time used where the frequency of 2100 MHz.

This paper presents the results of the quality of mobile Internet access based on LTE. The results of measurements of the downloading and uploading bitrates, the response times, the range and connection stability are presented. The measurements were done in Wroclaw by Kamil Żylewicz and the data for this paper were taken from his thesis [7]. As an Internet service provider has been selected P4 Company (Play operator), because it was the only one which not limited transmission speed of LTE, if was not sent more than 100 GB of data.

2 Organization of Measurements

The LTE connection quality with the Internet has been studied in:

- local home network in the center of Wroclaw, at Jana Kilinskiego str.,
- traffic on the A8 motorway bypass and the streets of Wroclaw.

In the home network measurements were performed every 10 min for two weeks, so that in total were obtained over 2 000 transmitted and the same number of received samples. During the experiment any updates have been excluded. The computer was also not used in any way, for example, to browse the web.

The measuring position was located directly next to the window, where GSM operator declares good LTE coverage inside of buildings.. The tests were done using Ookla Speedtest tool, which allows to analyze parameters of the data stream provided by the GSM operator to the user. For home network use were monitored: downloading and uploading bitrates, request response time.

For measurements in traffic, designated travel route shown in Fig. 1 with a length of nearly 40 km, including the exact center of Wroclaw, settlements with skyscrapers, single-family houses, roads in housing estates, main streets and Wroclaw Motorway Bypass A8. This enabled it to examine the connection quality in almost all conditions, at a speed of approximately 130 km/h. On the entire route GSM operator declares good LTE coverage outside of buildings. The measurements were done in a continuous manner (one after another) at night to avoid falsification of the results by the network load.

For mobile use was monitored: downloading and uploading bitrate, request response time, movement speed and received signal power.

For measurements were used:

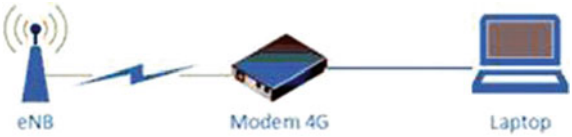
- belonging to the operator device eNB for communication with the modem,
- LTE modem Huawei E3272 with a transmission speed of 150 Mbps,
- TP-LINK TL-MR3020 router (only for measurements at home)
- laptop Samsung Ativ Book 6 (NP670Z5E-X01PL).

Fig. 1 Route of mobile measurements

Modem LTE Huawei E3272 configured to connect to the network only at LTE. Logic diagrams of measuring stations shown in Figs. 2 and 3.

Fig. 2 Measuring station in the home network

Fig. 3 Station for mobile measuring



3 The Results of Research in a Home Network

Despite locate measuring station in the city center, the signal power was not greater than 80–87 %. This could be due to the fact that around of the point of measurement on each side were quite high buildings (at least five floors) but the measurements were performed on the second floor. The signal could be significant dispersed as a result of reflections from the walls of buildings.

3.1 Bitrate

The measured bitrate of 2017 samples are within the ranges of 1.85 to 17.58 Mbps for downloading and from 0.54 to 12.0 Mbps for uploading data, as shown in Fig. 4 and in Table 1.

Figure 5 shows the download and upload bitrates averaged at certain times of the day within 14 days. The download bitrate was greatest between 3:00 am and 8:00 am and was the smallest between the hours of 5:00 to 10:00 pm. In the case of upload bitrate, it remained almost constant, increasing slightly in the hours from 3:00 to 8:00 am. Upload bitrate was the lowest from 1:00 pm to 10:00 pm, as shown in the graph in Fig. 5.

Most of the time (up to 94 %) download bitrate ranged from 5 to 15 Mbps. More often, because in 52 % of cases, the speed was greater than 10 Mbps.

When sending the data, in 68 % of samples the upload bitrate have ranged from 5 to 10 Mbps, but in 29 % the bitrate was less or equal to 5 Mbps. The highest bitrates exceeding 10 Mbps happened very rarely—in 3 % of all samples. The percentage of samples having a specific bitrates are illustrated in Fig. 6.

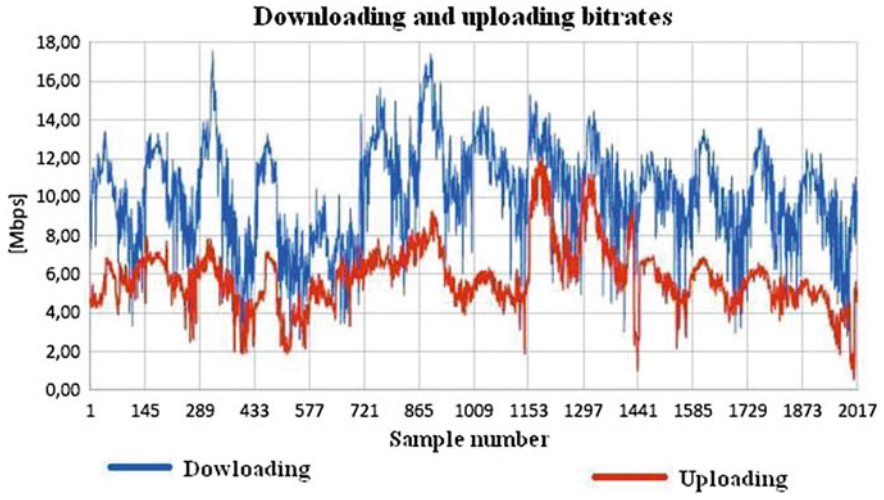


Fig. 4 Downloading and uploading bitrates for home network

Table 1 Downloading and uploading bitrates for home network

	Download [Mbps]	Upload [Mbps]
Minimum	1.85	0.54
Maximum	17.58	12.00
Average	10.06	5.83
Median	10.31	5.75
Standard deviation	2.61	1.66

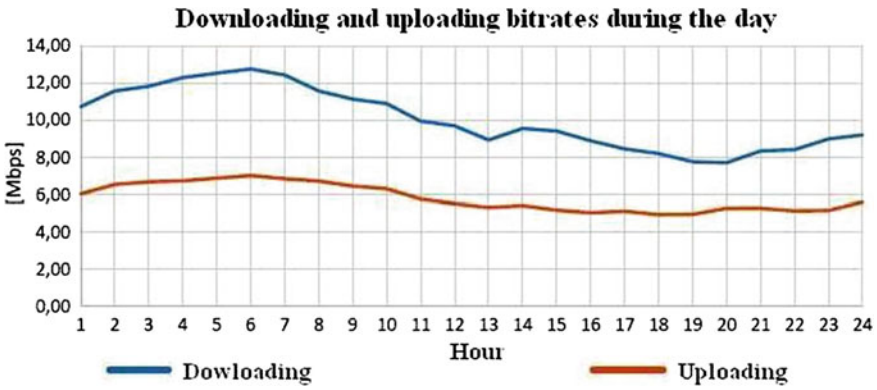


Fig. 5 Downloading and uploading bitrates during the day

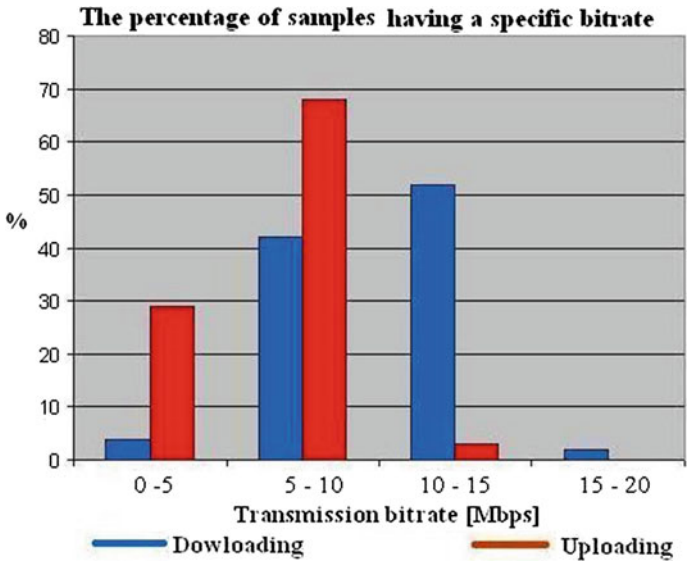


Fig. 6 The percentage of samples having a specific bitrates

3.2 Request Response Time

The time between sending the request and receiving the answers measured 2,017 times. During the research, before each measurement was elected the best server at the time, to avoid falsification of the results by the load of the target server. Measured values ware mainly influenced by the operator infrastructure and by the technical capabilities of LTE technology. The measurement results are shown in Table 2.

In terrestrial telecommunications networks delays up to 50 ms are accepted. The measured delay times deviate significantly from the declared specifications LTE, according to it the response time should be 5 to 10 ms, and such a result was not be achieved.

There is a clear relationship between response time and downloading and uploading bitrates. Higher response times correspond to lower bitrates. The correlation coefficient between response time and a downloading bitrate is −0.70, and between response time and a uploading bitrate is −0.72.

Figure 7 shows the linear regression of a standardized transmission rate as a function of a standard response time.

All three collections: downloading bitrates, uploading bitrates and response times were normalized to the range of values from 0 to 1, so that 0 was assigned to the minimum and 1 to the maximum according to the formula

Table 2 Response times measured in the home network

	Response times [Mbps]
Minimum	23.00
Maximum	171.00
Average	38.67
Median	34.93
Standard deviation	14.39

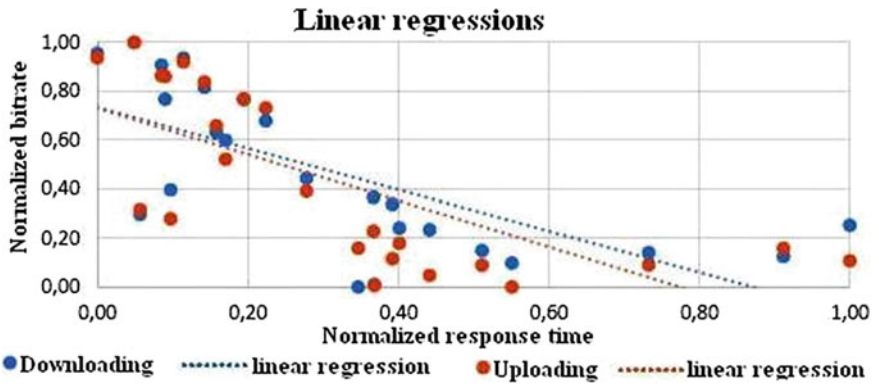


Fig. 7 Linear regressions of bitrates in dependence on the response times

$$X_i = \frac{x_i - \min(x)}{\max(x) - \min(x)},$$

where

X_i is the normalized value of value x_i,
$\min(x)$ is the minimum value of all x_i,
$\max(x)$ is the maximum value of all x_i.

4 The Results of the Measurements Carried Out in Motion

The biggest advantage of Internet access in LTE technology is its mobility. When approx. 1 h of measurements carried out in motion obtained 36 samples. Each sample describes the transfer bitrates in both directions, the response time, speed of movement, the current location and the signal power in percentage.

4.1 Bitrate

The bitrates of mobile downloading and uploading for 36 samples shown in Fig. 8. Measured bitrates shown in the graph in Fig. 8. In one sample achieved a record high download speed—more than 67 Mbps, which is shown in Table 3. This value close to the maximum declared by the operator, which is 70 Mbps. As we see in Table 3 the average downloading bitrate is close to 26 Mbps, while uploading bitrate is approximately 17 Mbps. The lowest measured speed in both directions

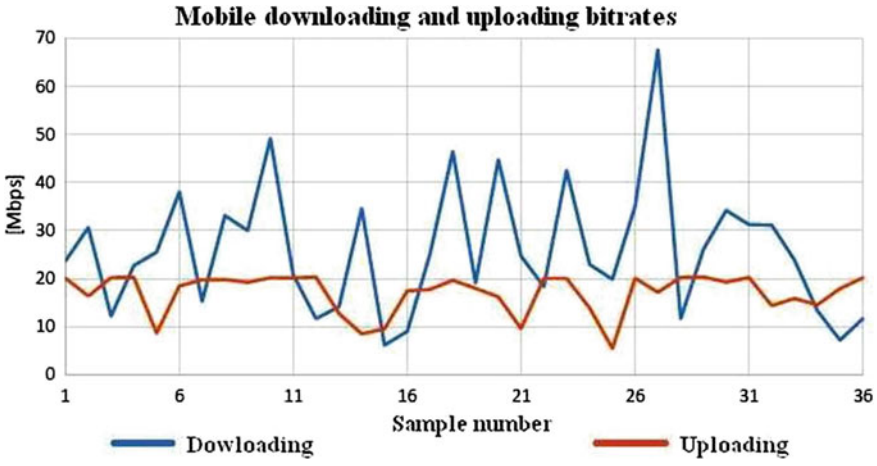


Fig. 8 Mobile downloading and uploading bitrates

Table 3 Mobile downloading and uploading bitrates

	Download [Mbps]	Upload [Mbps]
Minimum	6.14	5.44
Maximum	67.59	20.33
Average	25.92	16.99
Median	24.24	18.82
Standard deviation	13.39	4.13

was about 6 Mbps. Uploading bitrate proved to be more stable than the downloading bitrate. The standard deviation for uploading bitrate is more than three times larger than for downloading bitrate.

Downloading bitrate has changed significantly in a short period of time. This was due to the mobility of the measuring station and the base GSM stations. The impact of network load was small because the measurements were carried out late at night. The highest bit rate in excess of 40 Mbps, was observed in only 5 attempts.

As we see in Fig. 9, in over 91 % of the samples the downloading bitrate was satisfactory, it was greater than 10 Mbps. Simultaneously in almost 64 % of the time, the download bitrate was greater than 20 Mbps. The uploading bitrate in 75 % of samples exceeded 15 Mbps, including in 33 % of samples it slightly exceed 20 Mbps. Declines bitrates below 10 Mbps occurred in 14 % of the measurement time.

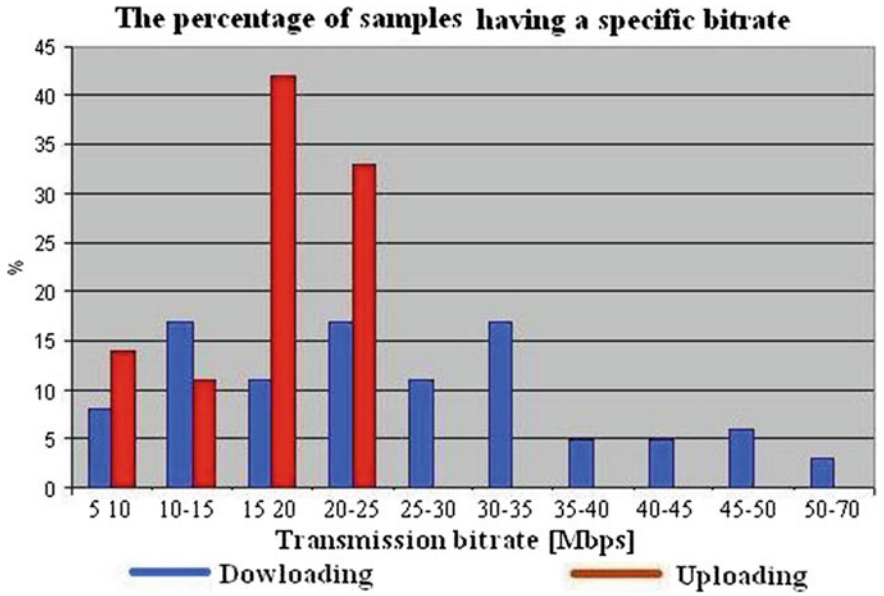


Fig. 9 The percentage of samples having a specific bitrate

4.2 Request Response Time

Request response time were measured 36 times in motion. In 72 % of all measurements the response time was from 24 to 30 ms. As we can see in Table 4, this were the lowest measured values. In 6 % of results the response time was from 31 to 40 ms, in 19 % it was from 41 to 50 ms, and in one case it was over 50 ms. The minimum, maximum and average values of the response time and its standard deviation are shown in Table 4. The download and upload bitrates and response times where measured at speeds up to 130 km/h. There was no significant effect of speed on the measured values.

4.3 Signal Power

When measuring in traffic also the signal power was measured. The measured values as a percentage of full power are illustrated in the graph in Fig. 10.

Table 4 Response times measured while driving

	Response times [Mbps]
Minimum	24.00
Maximum	65.00
Average	33.92
Median	29.00
Standard deviation	8.95

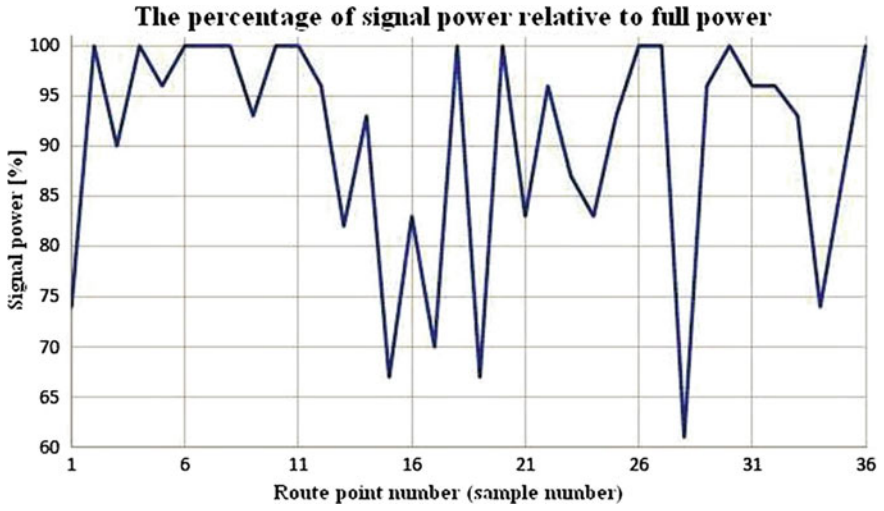


Fig. 10 The percentage of signal power relative to full power

Table 5 The percentage of signal power relative to full power

	Response times [Mbps]
Minimum	61.00
Maximum	100.00
Average	90.44
Median	96.00
Standard deviation	11.45

Power varied from 61 to 100 %, with an average of slightly more than 90 %, as detailed in Table 5. It can be considered that the signal level on a route was very good and coincided with declarations of operator. The best and most stable signal (samples from 2 to 11) was in the streets Armii Krajowej, Wisniowa and Jozefa Hallera. The lowest signal power (samples 15 to 19) was registered in the western part of Wroclaw.

The average signal power was at 90.44 %, and the median was 96 %, as are presented in Table 5. In 64 % of measurements obtained almost full signal power, exceeding 90 %, and the connection was not interrupted, which should satisfy even the demanding user.

5 Conclusion

The aim of this work was to investigate the quality of mobile access to the Internet implemented in the Long Term Evolution technology. In order to obtain large amounts of reliable data on the basis of which it was possible to draw the reliable conclusions, the investigation was carried out during of more than one month. At that time, there was sent nearly 200 gigabytes of data.

During measurements on a home network, the signal power was only 80–87 %, so the measured data transfer speed was smaller, and the response times were greater than expected. The average downloading bitrate which slightly exceeds the 10 Mbps, today is insufficient. The average upload bitrate close to 6 Mbps is acceptable and complies with offers of many operators. The use of LTE technology for home networks, as an alternative to cable connections to the Internet, can be justified.

LTE technology is ideal for the application for which it was designed, i.e. during movement. The entire length of the route traced in Wroclaw and on the motorway bypass, the connection to the GSM operator was active and was not interrupted. Transmission parameters significantly changed with the change of location. Movement speed (up to 130 km/h) does not have a major impact on the quality of the connection.

Acknowledgments This work was supported by statutory funds of the Department of Systems and Computer Networks, Wroclaw University of Science and Technology, grant S50020.

References

1. Dahlman, E., Parkvall, S., Sköld, J., Beming, P.: 3G Evolution—HSPA and LTE for Mobile Broadband, 1st edn. Academic Press (2007)
2. Dahlman, E., Parkvall, S., Sköld, J.: 4G—LTE/LTE-Advanced for Mobile Broadband, 2nd edn. Academic Press (2014)
3. Motorola. Long Term Evolution (LTE): A Technical Overview (technical white paper) (2007). www.3g4g.co.uk/Lte/LTE_WP_0706_Motorola.pdf
4. 4G Americas. 4G Mobile Broadband Evolution: 3GPP Release 11 & Release 12 and Beyond. www.4gamericas.org/files/2614/0758/7473/4G_Mobile_Broadband_Evolution_Rel-11__Rel_12_and_Beyond_Feb_2014__FINAL_v2.pdf (Feb 2014)
5. 4G Americas. Mobile Broadband Evolution Towards 5G: Rel-12 & Rel-13 and Beyond (white paper). www.4gamericas.org/files/6214/3569/1603/4G_Americas_Mobile_Broadband_Evolution_Toward_5G-Rel-12_Rel-13_June_2015.pdf (June 2015)
6. Play: LTE coverage in the Warsaw metro (in Polish). www.telepolis.pl/wiadomosci/play-zasieg-lte-w-warszawskim-metrze,2,3,32215.html (Dec 2014)
7. Żylewicz, K.: Analysis of the Services Quality of Mobile Internet access Based on Modern Technologies—Master's thesis (in Polish). Wroclaw University of Technology (2015)

SelfAid Network—a P2P Matchmaking Service

Michał Boroń, Jerzy Brzeziński and Anna Kobusińska

Abstract In this paper we propose an approach for discovering unused resources and utilizing them in on-line multiplayers games in P2P environments. The proposed SELFAID NETWORK automatically deploys and manages services running on machines belonging to end-users and connects players using the discovered resources. SELFAID NETWORK consumes only spare resources, following the trend of sharing economy.

Keywords Peer-to-peer · Unused resources · Nodes matchmaking · Sharing economy

1 Introduction

On-line multiplayer games attract a lot of attention nowadays. Developers of such games are not usually trained in networking, yet need to make their game work over the Internet. Popular game engines [3, 12] make it easy, by providing high-level features such as distributed object management or state synchronization. While these solutions perfectly abstract networking concerns, they rely on a client-server model, which implies additional costs for players. Whether running on a private infrastructure or in the cloud, servers generate maintenance costs. The price is ultimately paid by players in monthly fees or watching advertisements. In addition, this model makes it very difficult to release free multiplayer games. The recently popular idea of *sharing economy* may inspire an alternative model, free of the aforementioned flaws. Sharing economy relies on making use of un- or under- used resources. In the case of

M. Boroń · J. Brzeziński · A. Kobusińska (✉)
Institute of Computing Science, Poznań University of Technology, Poznań, Poland
e-mail: akobusinska@cs.put.poznan.pl; Anna.Kobusinska@cs.put.poznan.pl

M. Boroń
e-mail: mboron@cs.put.poznan.pl

J. Brzeziński
e-mail: jbrzezinski@cs.put.poznan.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_16

multiplayer games, those resources consist mainly of user's bandwidth and computing power. They are not utilized completely when a game is played at home, except for some extremely resource-hungry games. Sharing bandwidth with others does not impose additional costs, as most Internet Service Providers (ISPs) offer unlimited data plans to domestic users. Thus, employing this alternative approach would make it possible to develop multiplayer games which charge players only for game content.

The goal of this paper is to contribute a piece to the P2P gaming puzzle by providing a solution to the problem of matchmaking players on the scale of the Internet, without using a central server. Matchmaking in multiplayer video games is the process of connecting players together for online play sessions. The primary function of a matchmaking system is to find another person willing to play. Furthermore, a matchmaking system may contain features such as connecting players with similar abilities or with small latency. The proposed solution leverages unused resources belonging to players (specifically bandwidth and CPU) during the matchmaking process. Since there are many kinds of matchmaking strategies, implementing all of them in a single library would be a daunting task. Instead, it was noted that different matchmaking strategies may be thought of as separate server programs, which provide some *service* to other nodes. Consequently, the paper introduces the concept of a tool, called SELFAID NETWORK, which aims to locate and manage services in a P2P environment. As an example of using the SELFAID NETWORK, a matchmaking service connecting players based on a ranking system is presented. SELFAID uses only the machines belonging to end-users, so no additional infrastructure has to be deployed. Since using spare resources of just one node is not enough to fulfill the needs of popular computing games, the burden of providing a service is shared with others when it exceeds capabilities of a single node. Nodes receive only as much work as they can process using spare resources, and functioning of the SELFAID NETWORK does not interfere with other operations the node performs (such as downloading files).

The paper is organized as follows. In Sect. 2 we discuss the work related to ours. Section 3 contains description of the assumed system model. Section 4 presents the general idea of SELFAID NETWORK. Section 5 discusses the architecture and technical project of the proposed approach. Finally, Sect. 6 concludes the paper.

2 Related Work

This paper describes several works related to P2P multiplayer games, as well as their important shortcomings. The paper [9] proposed a distributed shared objects abstraction. However, the author didn't take issues of trust and safety into account and a central directory service had to be used in order to find other players. In [8] the synchronizing mechanisms for P2P Massively Multiplayer Games were offered. They assumed a static division of the game world and a coordinator for each game region. In the succeeding solution [13], an improved cooperation (provided incentive for

taking actions which benefit the whole community) in P2P systems by using social networks was proposed. It risks flooding the network with a lot of packets due to the nature of consignable requests mechanism. In turn, [15] showed how to make P2P games of turns secure. Unfortunately, the authors put a severe limitation on the size of playing group and didn't discuss how players find opponents. Solution described in [14], presented a design for enforcing security in P2P MMGs. It generates a lot of network traffic, because each action is broadcasted to all players in the region and a byzantine voting is frequently performed. Finally, [1] provided an efficient way to maintain consistent state between servers, but it doesn't discuss failure of matrix server and the simple rule of dividing the region in half may not be the most appropriate in some cases.

The majority of reviewed articles presented solutions for either synchronizing the game state [1, 8, 9] or contributed to other elements directly associated with playing the game [14, 15]. However, there is a research subject other than those two, which becomes visible after taking a closer look on the most important shortcomings presented above—[9] had to use a central directory service in order to find other players and [15] didn't discuss how players find opponents. Both works left out the problem of matchmaking players in a P2P environment. This paper aims to bridge the gap by providing a solution for this problem.

3 System Model

In this paper, we consider a peer to peer system model [2, 10]. The term P2P or peer-to-peer refers to direct communication between parties with equal rights. Some people incorrectly assume that fulfilling this criterion is enough for any system to be called P2P. What really brings together systems called P2P are the goals they aim to achieve and benefits associated with them. One of the goals is shifting the balance of computation from central servers to regular, personal computers. Another goal is to use otherwise unused distributed resources such as computing power, storage, network bandwidth. Common benefits include scalability and eliminating the need for expensive infrastructure such as servers and specialized networks.

The SELFAID NETWORK described in this paper is a decentralized and structured P2P network. All nodes in the system have equal roles—they perform the same tasks and have the same rights. As all structured P2P networks, the SelfAid network has a precisely defined set of rules on how nodes should choose with whom they establish and maintain connection. This set of rules is called a Distributed Hash Table [5, 11] and allows efficient resource location. Following the trend of sharing economy, SelfAid network assigns only as much work to a particular node as it can handle using spare resources.

SELFAID NETWORK automatically manages services running on machines belonging to end-users. Managing services consists of creating and destroying service instances, so that at any point in time it is possible to find and use the desired service. Furthermore, the proposed solution provides a way to locate nodes running

instances of a particular service, based on a unique service identifier. SELFAID consumes only spare resources of nodes: bandwidth and CPU power, following the trend of sharing economy. To connect to SELFAID NETWORK, it is sufficient to know an IP address of one of the nodes already present in the network.

SELFAID NETWORK operates under several assumptions. It assumes a failure-free environment that provides perfect links (messages are always delivered) and FIFO channels. Next, services are required to be stateless (data is not transferred between service instances). It is also assumed that service identifiers are known by the clients a priori. SELFAID NETWORK is responsible only for instantiating and providing contact information to reach service instances, so clients have to know the appropriate protocol to communicate with the service.

4 General Idea

The SelfAid network is a tool for automatic management of services in P2P environment, which can be used to build distributed matchmaking systems. In this section the concept of SelfAid network is explained and an idea for ranking based matchmaking system is presented.

From users point of view, it is enough to lookup (request contact information for) the service. The lookup procedure takes care of checking if a running instance of the desired service exists. The contact information (an IP address) of each service is periodically published, and called an announcement. The announcements are stored in a Distributed Hash Table (DHT). In order to determine whether a running instance of desired service exists, the lookup procedure tries to retrieve appropriate announcement from the DHT and pick contact information of one of the replicas. If it finds an announcement, contact information of an instance is returned to the user. In order to ensure proper load balancing between instances, contact information is chosen randomly, with uniform distribution of probability. On the other hand, if no announcement was found, an instance of the desired service is created on the node, which issued the request. Then its contact information (IP address and port) is returned to the user. It is possible, that many nodes simultaneously tried to lookup a service which was not run by anybody and ended up starting a service instance not knowing about each other. In order to deal with this situation, the announcement contains a hash of the original node identifier, hash of service name and running time — the time which elapsed since the moment, in which the original node started its service instance. A node, which received request to store an announcement stores it if it contains a higher value of running time than the one currently stored. If its impossible to tell which service instance is running for a longer time, due to networking delays, the announcement with the hash value of the original node identifier closer to the service name hash value is stored. The nodes publishing announcements must periodically check if their announcements are stored. To prevent spreading outdated information, an announcement is stored for a limited amount of time and then deleted.

Instances of a particular service running on SELFAID NETWORK are organised in a ring structure. The number of running instances of a service depends on the demand for the service (bigger demand means more nodes). The first instance of a service is launched as the result of a failed attempt to lookup an announcement for the service. From that point forward, additional instances are automatically created to ensure service availability. When load on a service node becomes too big/too small, the node notifies the SELFAID NETWORK to adjust the number of instances. If there is a need to increase the number of service instances, one of the nodes of SELFAID NETWORK which is not running an instance of this service is asked to be a part of the ring—one of possible ways to reach such node is to send a message asking for cooperation to a random node of SELFAID. To shrink the number of instances, a node is asked to shutdown and the node watching it is notified.

The proposed matchmaking system connects players based on their ranking. There are multiple player ranking algorithms available such as ELO [4] or TrueSkill [6]. For the purposes of this article, a simplified ranking algorithm, similar to the system used in the game Hearthstone, is chosen. It is assumed that each player has a rank, represented by a number from the range of 1 to N, where N is the number of the lowest (the least skilled) rank. A player can connect only with players of the same rank (so that he plays only with people of similar skill). For each win, player is awarded a point and for each loss player loses a point. After accumulating three points the rank of the player increases and points are reset. If a player loses a game while having 0 points, his rank is decreased by 1.

The matchmaking system creates a separate service in SELFAID NETWORK for each rank. The matchmaking process looks as follows:

1. player locates the node providing matchmaking service for his rank
2. player asks service for candidates, sending his *blacklist*
3. service adds player to the set *WS* of waiting nodes
4. service sends candidates from (*WS**blacklist*)
5. player sends parallel requests to all candidates
6. if not in play, candidate measures latency to player
7. if latency is acceptable, candidate asks player to initiate play
8. if player is not yet in play, he confirms and the game starts
9. player and candidate notify the service that they found an opponent
10. service removes the player and candidate from *WS*

When a player wants to find a partner to play with, he looks up an announcement of a ring corresponding to his rank. Verifying the rank claimed by the player is outside of the scope of this article. This problem could be solved e.g. by using the solution proposed in [14] or [15]. Then, the player picks an address of a service node from the announcement and asks it for the list of candidates who could become his opponents. The player also sends a blacklist. The blacklist is a list of candidates, with whom the player already tried to connect with but failed for some reason. After receiving a message from the player, the service adds the player to a set of nodes waiting to be connected with others. Next, the service sends a list of candidates consisting of waiting nodes except for nodes blacklisted by the player. After receiving

the list of candidates, the player tries to initiate latency measurement with all candidates. A candidate will agree to measure latency if it is not already playing the game. If the measured latency is acceptable by both the player and the candidate and none of them is already playing, they agree to play together. In case all candidates refuse playing together, the player goes to step number 1. Both player and candidate notify the service node that they found an opponent. Then the service node removes them from the *WS*, so that they don't receive unneeded messages.

5 Architecture of SELFAID NETWORK

The architecture of the proposed solution is shown in Fig. 1. The project is composed of five main modules.

`Kademlia` module is a modified implementation of Kademlia DHT by [7]. It leaves the storage implementation to the user and notifies the storage of `PUT` and `GET` requests using callbacks. `Announcement` and `GetFreeNode` modules contain implementations of announcement mechanism and adding nodes to the SELFAID NETWORK, described in Sect. 3. Both of them use `Kademlia` module and define a custom storage implementation. `SelfAid` module contains message processing routines (serializing, deserializing, queuing, removing duplicates, etc.), implementation of algorithm managing instances, sketched in Sect. 3, and classes meant to be used by a user of `SelfAid` module. `Matchmaking` module uses functionalities provided by SELFAID to create a simple distributed matchmaking system.

There are no restrictions imposed on how an object to be stored in `Kademlia` module should be defined. The only requirement is that the class extending Kademlia storage has to be able to tell if it is storing an object for a given Kademlia identifier (`KademliaId` is a class representing an identifier of a Kademlia node or hash of the sought value). Three types of messages are implemented in the context of this module: store (`DHT.PUT(...)` request), lookup (`DHT.GET(...)` request), found (`DHT.GET(...)` response). The stored and found messages have to contain the user-defined object to be stored/which was retrieved. Lookup message has to be

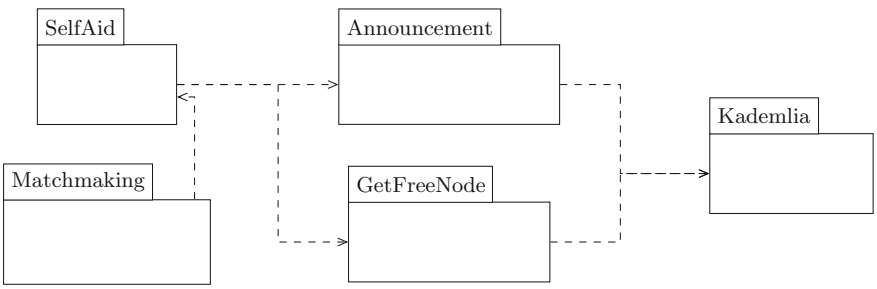


Fig. 1 Dependencies between main modules

defined for Kademlia to know which type of content is looked up in case there are multiple storages defined. All three messages have to contain a unique identifier. A user has to provide a storage mechanism by extending the `KademliaStorage` interface. Methods of this interface are called by a thread in `Kademlia` module when `DHT.GET` or `DHT.PUT` request comes. When a request to get a content of Kademlia object is obtained, such a content is returned (assuming it was found).

`SelfAid` module serializes (deserializes) the message to (from) output (input) stream, and performs the action associated with receiving a message, which specifies how the system handles it. Since reaction is tied to message implementation, it has access to all fields of the message. It is also responsible for sending or broadcasting messages. It maintains outgoing clocks and resends messages (for reliability) automatically. Internally, it operates on sockets: sends packets and listens for incoming packets. It maintains incoming clocks and buffers messages automatically. Received messages are added to a FIFO queue. When a message is sent, its code is written to the packet before message content. The receiving node is able to deserialize message correctly—the idea is to use the abstract factory pattern and automatically (based on received message code) choose appropriate factory to create instance of concrete message. Nor message receiver, nor any other component is expecting to know what kind of messages it processes. All of them use only methods defined by the abstract Message class.

User has to create an implementation of `ServiceManager`, which will enable `SelfAid` module to start and shutdown instances of user-defined services. The implementation of `ServiceManager` receives a `LoadReporter` object in the call to launch method. When load becomes too big/too small/ok, the implementation of `ServiceManager` calls appropriate method on the received `LoadReporter` instance. The implementation has to return the port number of launched server, so that SELFAID can properly announce its presence. To properly start SELFAID NETWORK the following steps have to be taken. User creates one instance for each (his) implementation of `ServiceManager`. Then for each created instance, user has to instantiate a `ServiceManagerRegistryEntry`, which is the name of the service (it has to be unique, later a `KademliaId` key is computed based on this value). Next, user creates a `ServiceManager` and adds to it all created entries. Finally, a `SelfAid` object can be created. User passes the registry instance as an argument to `SelfAid` constructor. Later, user calls lookup Service method on SELFAID with the name of service, which he wants to access. It is also possible to gracefully shutdown, by calling shutdown method.

`Matchmaking` module contains the logic of ranking based matchmaking, which is built on top of the SelfAid abstraction. The logic operates as described in "System model and general idea". The module defines a `MatchmakingClient`, which sends a request to the matchmaking server and awaits response with candidate list. Then, it follows the steps described in the previous section. `Matchmaking Service` is the implementation of server part. It handles gathering the addresses of players seeking opponents and responding to players with candidate lists, as well as measuring load on the server. Load is measured by putting an entry to a `HashMap` every time, a packet is received. The entry contains time of receiving the packet

(in milliseconds). A thread periodically traverses the map and calculates the number of entries which were inserted maximum 6 seconds before (all other entries are removed). Based on the number of received packets during this timeframe, load may be reported as too big/too small or ok using a reference to `LoadReporter` class from `SelfAid` module. `MatchmakingServiceManager` is an implementation of SELFAID `ServiceManager`, which creates or shuts down an `Matchmaking Service` instance on demand of SELFAID.

6 Conclusions and Future Work

This paper provided the general idea and initial description of the architecture of the SELFAID NETWORK tool for automatic management of services instances in P2P environment. The proposed tool was usesd to build a distributed matchmaking system for on-line game players, which uses only spare bandwith and CPU nodes resources. All nodes in the proposed system may perform any role, depending on circumstances. The system is able to connect as much nodes as the underlying Kademlia DHT, which was designed for massive scale. Due to the modular architecture, it is possible to easily extend the system with new functionalities. The proposed solution could be combined with [9] to create a completely distributed online gaming framework.

The authors plan to enhance the proposed solution. Primarily, the current version assumes that no failures occur. Thus, in the next step we are going to consider the environment susceptible to failures. Furthermore, creating a simulation with a big number of nodes is planned, to analyze the performance and scalability of the system in depth.

References

1. Balan, R.K., Ebling, M., Castro, P., Misra, A.: Matrix: Adaptive middleware for distributed multiplayer games. In: Middleware 2005, ACM/IFIP/USENIX, 6th International Middleware Conference, Grenoble, France, November 28–December 2, 2005, Proceedings. Lecture Notes in Computer Science, vol. 3790, pp. 390–400. Springer (2005)
2. Barkai, D.: Peer-to-Peer Computing: technologies for sharing and collaborating on the net. Intel Press (2001)
3. Developers, G.: Google Cloud Platform dedicated server gaming solution. https://cloud.google.com/solutions/gaming/dedicated-server-gaming-solution/
4. Elo, A.E.: The Rating of Chessplayers, Past and Present. Arco Pub, New York (1978)
5. Galuba, W., Girdzijauskas, S.: Distributed hash table. In: Liu, L., A-zsu, M.T. (eds.) Encyclopedia of Database Systems, pp. 903–904. Springer, US (2009)
6. Herbrich, R., Minka, T., Graepel, T.: Trueskill^{tm}: A bayesian skill rating system. In: Schölkopf, B., Platt, J.C., Hoffman, T. (eds.) Advances in Neural Information Processing Systems 19 (NIPS-06). pp. 569–576. MIT Press (2007)
7. Kissoon, J.: Joshuakissoon/kademlia. https://github.com/JoshuaKissoon/Kademlia

8. Lu, H., Knutsson, B., Delap, M., Fiore, J., Wu, B.: The design of synchronization mechanisms for peer-to-peer massively multiplayer games. Department of Computer and Information Science, The University of Pennsylvania Technical Report (2004)
9. Ørbekk, K.: Distributed shared objects for mobile multiplayer games and applications. In: Master Thesis. Institutt for datateknikk og informasjonsvitenskap (2012)
10. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: a scalable peer-to-peer lookup service for internet applications. ACM SIGCOMM Comput. Commun. Rev. **31**(4), 149–160 (2001)
11. Stoica, I., Morris, R., Liben-Nowell, D., Karger, D.R., Kaashoek, M.F., Dabek, F., Balakrishnan, H.: Chord: a scalable peer-to-peer lookup protocol for internet applications. IEEE/ACM Trans. Netw. **11**(1), 17–32 (2003)
12. Unity3D: Unity3D description. https://unity3d.com/
13. Wang, W., Zhao, L., Yuan, R.: Improving cooperation in peer-to-peer systems using social networks. In: Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International. pp. 50–57. IEEE (2006)
14. Wierzbicki, A.: Trust enforcement in peer-to-peer massive multi-player online games. In: On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE, pp. 1163–1180. Springer (2006)
15. Wierzbicki, A., Kucharski, T.: Fair and scalable peer-to-peer games of turns. In: Proceedings of 11th International Conference on Parallel and Distributed Systems, 2005, vol. 1, pp. 250–256. IEEE (2005)

Ant Colony Optimization in Hadoop Ecosystem

Marek Kopel

Abstract Paper focuses on bringing the classic ACO (Ant Colony Optimization) for TSP (Travelling Salesman Problem) to Hadoop ecosystem. Classic ACO can be parallelized for efficiency. Especially today, with virtualization and cloud computing it is particularly easy to run ACO simulation on many nodes. However the distribution part adds an extra cost to an implementation of a simulation.

Keywords Hadoop · MapReduce · Ant colony optimization · Travelling salesman problem · ACO parallel implementations

1 Background

The ant colony optimization (ACO) is a nature inspired probabilistic technique for solving computational problems. Since the algorithm is inspired by the work of ants, problem to be solved by ACO should be reduced to finding specific paths through graphs. This algorithm is a member of the a larger family of swarm intelligence methods. According to Marco Dorigo, who first introduced the ACO concept in [3], as he has written in [4]: "Since the proposal of the first ACO algorithms in 1991, the field of ACO has attracted a large number of researchers and nowadays a large number of research results of both experimental and theoretical nature exist. By now ACO is a well established metaheuristic".

The most popular problem used with ACO is Traveling Salesman Problem. The TSP is formulated by a question: What is the shortest path for visiting all the given cities only once and come back to starting point, assuming the distances between the cities are known? TSP is a NP-hard class problem. This is why using a metaheuristic like ACO is usually best approach.

M. Kopel (✉)
Wrocław University of Science and Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: marek.kopel@pwr.edu.p
URL: http://www.ii.pwr.wroc.pl/∼kopel

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_17

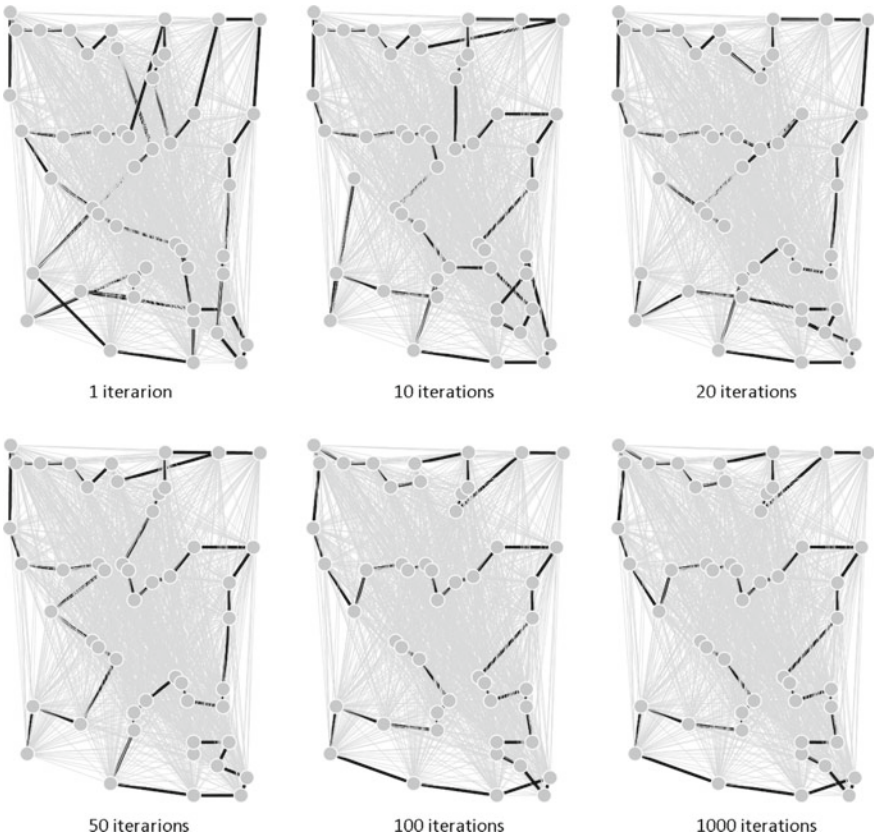


Fig. 1 Each graph shows the shortest path found after corresponding number of iterations of ACO run on a 50 node TSP graph (for readability the returning edge is not drawn)

ACO method tries to emulate the way ants find the best (shortest) path between their nest and food by placing pheromone. Ants succeed because of the scale effect. The cooperation among multiple ants is what ACO implementations try to use solving e.g. TSP, by placing food in graph nodes, which represent salesman target cities. And since each emulated ant within a programmed colony operates virtually the same way, there is a great potential for running ant algorithms in parallel.

Standard ACO run have all colony ants travel through food nodes and come back to first, nest node. Best solution is taken from the ant that took the shortest path. Single run cannot give great results, so runs have multiples iterations. In Fig. 1 best solutions are presented after various number of ACO iterations.

The 6 TSP graphs with best solutions demonstrate a serious problem in ACO implementations. The graphs are plotted using d3js[1] force layout, which tries to preserve edge weights as fixed-distance geometric constraints. After a few iterations the

[1] https://d3js.org/.

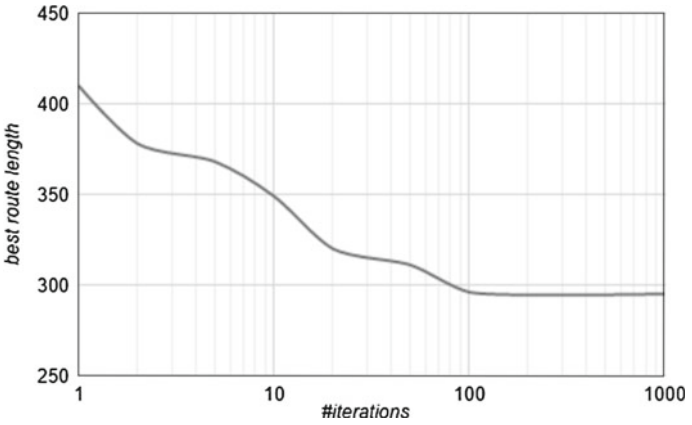


Fig. 2 Best route length function quickly flattens out. N.B. number of iterations is on a logarithmic scale

path is far from optimal: it goes from one borderline node to another, often crosses itself. Then after 50 and 100 iterations the path is quickly improved, but then after another 900 iterations there's hardly any change to the solution. This demonstration shows problematic characteristic of ACO: the solution gets major improvements in relatively short time, but then the improvement rate slows down to the point where there's almost no improvement with next iterations. This can be seen in Fig. 2. Half the the improvement was found in first 10 iterations. And the other half in another 1000 iterations. Cruoial decision needed to be taken in every ACO implementation is: when to stop the iterations, because the cost is not worth the improvement.

The problem comes from the fact that the function in Fig. 2 usually is not monotonically decreasing and it is impossible to foresee when the next big improvement in best solution may come.

To solve this problem whole ACOs are run in parallel to maximize the chance of not falling into a local minimum and obtaining the closest to optimal solution with fewer iterations.

2 Related Works

According to [9] the parallelization of ACO may be measured using 2 metrics: speedup, which is the time saved when using a parallel algorithm version instead of a sequential one; and efficiency, which is a normalized speed up to compare different hardware configurations. Authors use those metrics to introduce the state-of-the-art taxonomy for parallel ACO algorithms leaving out the multiobjective and dynamic versions of the problem.

Historically there have been some concepts for high-level categorizations of parallel ACO approaches. Authors of [10] distinguish operation models based on master-slave paradigm, independent executions and synchronous cooperation. In [6] algorithms are divided to: standard and custom parallelization of ACO, with the last one trying to gain efficiency from the specifically designed algorithm behavior. Introducing the new taxonomy in [9], authors reuse and expand concepts from those two works. Eventually the taxonomy includes four main model categories and a fifth: '*hybrid models*' for proposals that would fit to more than one main category. The four main categories are:

- '*master-slave model*', where multiple clients are controlled in a centralized way by a server;
- '*cellular model*', in which small, overlapping neighbourhoods in a single colony, each solving its part of the problem;
- '*parallel independent runs model*', where ACOs are executed concurrently, without any communication among them; and
- '*multicolony model*', for multiple ACOs interacting with each other by periodically exchanging best solutions.

The four categories may be grouped in pairs, but different pairs depending on the aggregation criteria. When dealing with a single colony: ACO may be *master-slave* or *cellular*. And for multiple colonies the models are: *parallel independent runs* and *multicolony*. The other criteria is the cooperation The only couple of models in which different parts work on a common solution is *cellular* and *multicolony*. More granularity than the 'four main plus one' categories was only proposed for the *master-slave model*—the most popular one. Depending on the level of interaction between server and clients there are three subcategories:

- '*coarse grain master-slave*', where only complete solutions are exchanges;
- '*medium grain master-slave*', where the problem is decomposed among the clients; and
- '*fine grain master-slave*', with the highest specialization of the clients and interaction level.

The most popular model of ACO parallelization is *master-slave* in its *coarse grain* version. An example of that approach is presented in [1], where authors—using an 8 CPU cluster—demonstrate how big speedup can be achieved by increasing number of ants at each processor.

The multicolony approach can be found for instance in [7]. The ACOs here are run on a homogeneous cluster of 4 computational nodes running Linux and LAM/MPI as communication libraries. The inter-ACOs cooperation was tested with a spectrum of 5 configurations with less and less communication configured: from 'fully-connected' to 'parallel independent runs'.

The approach researched in this paper can be thought of as *parallel independent runs*. Its classification is determined by Hadoop ecosystem characteristics. Figure 3 places this approach in the context of cited taxonomy.

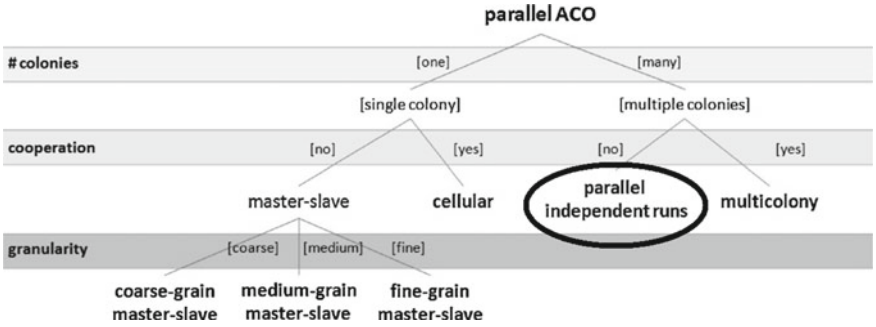


Fig. 3 A hierarchical view of parallel ACO taxonomy proposed in [9] with ellipse marking this paper's approach

3 Hadoop Ecosystem

Apache Hadoop [5] is a software framework for distributed storage and distributed processing of—so called—*big data* on computer clusters. It became very popular recently for cost effective parallelization of data processing in many domains. But using this framework for ACO parallelization demands a little different approach than the classic ACO parallelization.

The storage part of the framework, called Hadoop Distributed File System (HDFS) allows taking advantage of *data locality*, which means that each node in a cluster processes the data with a local access. This is why within a single Hadoop job the input files are being split into large blocks and being distributed across the nodes to process. Then each block is being processed by a single task. The data locality approach constrains tasks from sharing the data with one another.

This is way Hadoop framework is not suited for fine grain ACO parallelization, since the grains here would be as 'coarse' as the input blocks for each node. In this case, each block is an ACO for a complete TSP solution. Anyway, TSP itself cannot be solved in a fine grain manner, since the whole distance matrix and the whole pheromone matrix are always needed. The Hadoop approach could work as a coarse grain master-slave model, but unlike this model—Hadoop can make use of multiple ACOs. And since Hadoop tasks are isolated, the sharing constraint makes it impossible to implement parallel multicolony in this framework.

The other part of Hadoop framework, beside HDFS, is the processing part called MapReduce. Implementing processing of the data in the framework means implementing 2 functions: *map*() and *reduce*(). *Map* dispatches the job data to task slots and *reduce* aggregates the task results.

4 Hadoop ACO Implementation

The idea for the Hadoop approach in this work is to use ACO code made available with [2]. Every ACO for TSP instance is run as a task within the map function. Then the reduce function aggregates the results for best solution (shortest path). It is similar approach to the implementation of ACO parallelization with Hadoop presented in [8]. It's hard to make an in-depth comparison of the two approaches since the authors of [8] give too little operational details.

The process of running parallel ACOs for TSP works as follows. All data needed for a TSP solution, i.e. distance matrix, number of ants and number of iterations is serialized to JSON object. The object is then placed as a single line of a plain text file. The number of JSON objects and thus the size of the file is dependent on the number of parallel ACOs to be run. After dumping the serialized input data, the file is placed in HDFS system, from where it can be used by MapReduce. Serialization and deserialization of JSON objects is done with Gson[2] library.

The Hadoop part starts by mapping each line of the input file—each containing a TSP problem—to a separate ACOs. After deserialization from JSON, each colony runs by processing the input. After a number of iterations specified in input an output object is created, serialized to JSON and passed to reducer. The reducer writes the serialized objects back to HDFS as a plain text output file and a single Hadoop process is finished. Then the output file is taken from HDFS to read the computed best paths.

5 Experiment

The experiment idea is to run sequentially the ACOs with the classic implementation and then compare time it took with time taken by Hadoop implementation. Each ACO has 50 nodes, 50 ants and 50 iterations. Each time the distance matrix and pheromone matrix are the same. So there is no sense to compare the actual results of the optimization—which is the shortest TSP path—since the code and all input data is the same. The only factor that differs the two approaches is the run time. The Hadoop benefiting from parallelization can reduce the time needed to run the ACOs in sequential manner.

The experiment was run on a standard Windows 10 desktop (Intel Core 2 Duo 2.66 GHz, 8 GB RAM) running a docker container with Hadoop configured on a CentOS Linux distribution. Run times of both implementations are presented in Table 1.

[2] https://github.com/google/gson.

Table 1 Time and speedup of running a corresponding number of ACO in sequential and parallel manner

	Sequential		Parallel	
#ACOs	Time (s)	Speedup	Time (s)	Speedup
1	5.4	1.00	13.9	1.00
2	17.5	0.61	16.5	1.68
4	42.1	0.51	20.6	2.70
10	94.1	0.57	36.0	3.86
20	218.2	0.49	53.2	5.23
50	592.6	0.45	98.9	7.03
100	1329.1	0.40	183.9	7.56
200	3011.5	0.36	342.4	8.12
500	8195.2	0.33	844.5	8.23
1000	16640.8	0.32	1685.5	8.25

6 Discussion

The experiment results support the underlying intuition: parallelizing the ACOs with Hadoop is possible. Even though MapReduce not is used exactly the way it was supposed to—since TSP is hardly a *big data* problem—using it for parallelization gives great benefit. An existing code was reused with no change within standard map and reduce functions. The docker container with preconfigured Hadoop environment was used with little effort. One may say that the parallelization was virtually costless. And the benefit much better than expected.

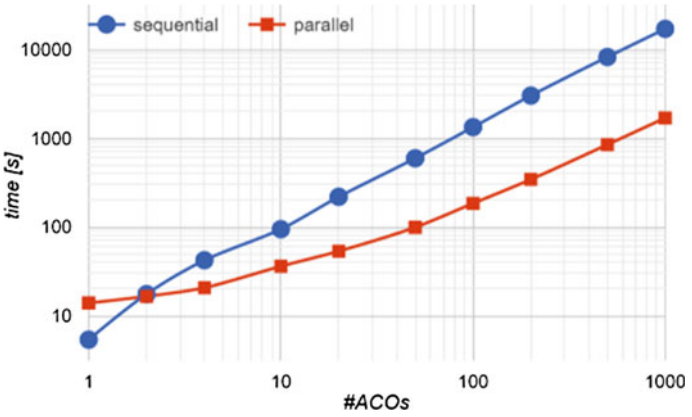


Fig. 4 Comparison of run times of multiple ACOs in two implementations: sequential (standard) and parallel (Hadoop) on both logarithmic scales

The Hadoop parallelization speedup was over 8 times before it started to flatten out. But what is unexpected the speedup of sequential runs is falling far below 1. It actually get even below 0.33 which may indicate ineffectiveness of Java runtime for long processes.

As can be seen on Fig. 4, time needed for running multiple ACOs grows much slower for Hadoop parallel approach than sequential runs. The sequential approach times show its non-linearity, which in theory should be linear, i.e. running ACO 2 times should take twice as much time; 3—three times as much time, etc. But since the sequential implementation behaves even worse than expected the overall parallelization benefit is even bigger.

7 Conclusions

This paper's goal is finding costs and benefits of using Hadoop framework to parallelize Ant Colony Optimization processes for solving Traveling Salesman Problem. The experiment of comparing run times of ACOs in standard and parallel manner showed the actual benefit of the latter. Also an unexpected outcome was the poor performance of the sequential ACO implementation. Running the experiment also proved the parallelization to be cost effective. Overall the experiment results show that using Hadoop for parallel ACO runs can only be beneficial.

Hadoop tasks are isolated so there can be no communication between any two parallel ACO tasks. But assuming the updated pheromone matrices can be preserved after Hadoop job is ended, they can be used again in next jobs. This way working in stages, similarly to approach proposed in [8], the Hadoop implementation can use collaboration among earlier and latter ACOs, and thus become more of the multi-colony model. This shall be further researched.

References

1. Chintalapati, J., Arvind, M., Priyanka, S., Mangala, N., Valadi, J.: Parallel ant-miner (pam) on high performance clusters. In: Swarm, Evolutionary, and Memetic Computing, pp. 270–277. Springer (2010)
2. Chirico, U.: A java framework for ant colony systems. In: Ants2004: Forth International Workshop on Ant Colony Optimization and Swarm Intelligence, Brussels (2004)
3. Dorigo, M.: Optimization, learning and natural algorithms. Ph.D. Thesis, Politecnico di Milano, Italy (1992)
4. Dorigo, M., Stützle, T.: Ant colony optimization: overview and recent advances. Techreport, IRIDIA, Universite Libre de Bruxelles (2009)
5. Hadoop, A.: Welcome to apache hadoop. http://hadoop.apache.org. Accessed 13 Feb 2016
6. Janson, S., Merkle, D., Middendorf, M.: 8 parallel ant colony algorithms. In: Parallel Metaheuristics: A New Class of Algorithms, vol. 47, p. 171 (2005)
7. Manfrin, M., Birattari, M., Stützle, T., Dorigo, M.: Parallel ant colony optimization for the traveling salesman problem. In: Ant Colony Optimization and Swarm Intelligence, pp. 224–234. Springer (2006)

8. Mohan, A., Remya, G.: A parallel implementation of ant colony optimization for tsp based on mapreduce framework. Int. J. Comput. Appl. **88**(8) (2014)
9. Pedemonte, M., Nesmachnow, S., Cancela, H.: A survey on parallel ant colony optimization. Appl. Soft Comput. **11**(8), 5181–5197 (2011)
10. Randall, M., Lewis, A.: A parallel implementation of ant colony optimization. J. Parallel Distrib. Comput. **62**(9), 1421–1432 (2002)

Measuring Efficiency of Ant Colony Communities

Andrzej Siemiński

Abstract The paper presents a study on the efficiency measures of the Ant Colony Communities (ACC). The ACC is an approach to parallelize the Ant Colony Optimization algorithm (ACO). An ACC is made up of a Community Server that coordinates the work of a set Ant Colony clients. Each client implements a classical ACO algorithm. The individual colonies work in an asynchronous manner processing data sent by server and sending back the obtained results. There are many possible locations for the clients: the same computer as the server, computers of a local or wide area network. The paper presents a detailed description of concept the ACC and reports the study of the efficiency of the such Communities. The efficiency is measured by their power (the amount of data processed in a given period of time) and scalability—the efficiency of adding colony clients on the Community. The paper contains also the taxonomy of parallel implementations of the Ant Colony.

Keywords Ant colony optimization · Travelling salesman problem · Parallel implementations · Sockets · Task complexity · ACC power · ACC scalability

1 Introduction

The aim of the paper is to present the Ant Colony Community (ACC) and different measures for the evaluation of its performance. The ACC is a set of Client Ant Colonies implementing the Ant Colony Optimization (ACO) metaheuristic algorithm and a Colony Server. The task of the Server is to send to its Clients cargos—packages of data to process. The server is also responsible for collecting and integrating partial results sent by client colonies. The main advantage of the ACC is

A. Siemiński (✉)
Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wrocław, Poland
e-mail: Andrzej.Sieminski@pwr.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_18

that is offers a parallel, asynchronous, easily scalable operation of many clients. This makes it possible to reduce the main disadvantages of the ACO metaheuristic —long processing time. The ACC is used to solve the Travelling Salesman Problem (TSP)—one of classical problems of Artificial Intelligence.

The TSP is widely regarded as a touchstone for many general heuristics devised for combinatorial optimization. This is a remarkably simple problem: given a list of cities and the distances separating them what is the shortest possible route that visits each city exactly once? The TSP simplicity of formulation is in a sharp contrast to its complexity. The number of all possible different routes for a graph with n nodes (cities) is equal to $(n - 1)!/2$ which is estimated by $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$. Even for relatively small values of n like 50 the value exceeds by far the mass of an observable steady-state universe expressed in kilograms. Solving the TSP is an active research area and a recent comparison of metaheuristics used for solving it could be found in [1]. In this paper we use Any Colony Optimization metaheuristic for that purpose.

The paper is organized as follows. The second Section introduces the ACO version for the Travelling Salesman Problem. In particular it discusses computational complexity of solving a task. The huge number of calculations that are necessary to complete an optimization process make parallel operation very attractive. Section 3 presents the taxonomy of parallel implementations of the ACO's concept. The main contribution of the paper—the Ant Colony community (ACC) is introduced in Sect. 4. It describes its' schema, basic operation principles and message passing rules. The low level criteria for the ACC evaluation are introduced in the Sect. 5. Section 6 deals with the ACC structure optimization. It describes an analytical tool to predict the power of an ACC and verifies its usefulness. The paper concludes with the resume of work done so far and indicates future research areas.

2 Analysis of the Basic ACO Operation

The basic ACO was proposed by Doringo [2]. There same variants of the basic ACO concept such as Ant System, MIN-MAX Ant System [3] but the main principle of work remains the same. The graph describing the locations of the cities or nodes is represented by a floating point two dimensional array with the distances separating the cities. Optimization process starts with randomly scattering the ants over the cities. An ant colony finds solutions in an iterative manner. In each one an ant travels from one city to another depositing a pheromone on its way. An iteration completes when the ants have visited all cities. At that very moment the pheromone levels of the whole matrix are updated. The colony remembers the Best-So-Far (BSF) route and its length. One of the key features of the ACO is its indeterminism. As a result the BSF route could occur at any iteration.

Selecting optimal values for the operation of the ACO is not an easy task [4, 5]. The studies reported in [6] point out that increasing the number of used ants could speed-up the convergence of results without increasing the processing time.

2.1 Estimating Optimization Task Computational Complexity

The operation of an ACO is time consuming. A good measure of the complexity of the optimization task is the number of calls of the quality function (*qf*) that calculates the usefulness of each node. It requires the calculation of the floating point power function. The power function requires much more time to execute than any other instruction e.g. on a regular computer the floating point multiplication takes just a few nanoseconds whereas the power function needs almost 1 microsecond to compute.

Let *nNodes* and $qfN(nNodes)$ denote the number of nodes in a graph and the number of calls to the *qf* function made by a single ant during one iteration. To complete an iteration an ant has to select nNodes-2: the first node is given to it at the start of an iteration and after making nNodes-2 selections the last node is obvious. The value of the $qfN(nNodes)$ is calculated buy the formula (1).

$$qfN(nNodes) = \sum_{k=1}^{k=nNodes-1} (nNodes - k) = \frac{nNodes(nNodes-1)}{2} - 1 \qquad (1)$$

For one iteration and the number of nodes equal to 50 a single ant needs 1224 calls of the *qf* function to complete its work. The function has to be calculated anew each time it is evoked because the pheromone array is shared by all ants and the pheromone levels are updated all the time. In a typical optimization process many dozens ants and hundreds of iterations are used. As a result many millions *qf* function calls are necessary to find solution. Dealing such large numbers is not convenient and therefore we introduce the *SC* constant equal to 3060000. The SC represents the number of calls of *qf* that are necessary to process the so called Standard Cargo task. The Standard Cargo optimization task consist of finding a solution for a matrix with 50 nodes using 50 ants and 50 iterations. The numbers were selected in such a manner as to make easy the comprehension of complexity of different optimization task. The processing the Standard Cargo by the most computers used in the experiments requires over one second. This also facilitates the evaluation of achieved results. The complexity of a task *tk* is measured by $Sc(tk)$ function. Its value is the number of calls to the *qf* function that are necessity to complete the optimization task *tk* expressed in the SC units. It depends on the number of ants, iterations and nodes.

Table 1 Complexity of optimization task

Ant # number	Iteration #	Node #	Complexity
50	50	50	1
50	1,000	50	20
100	1,000	50	40
150	500	50	30
100	1,000	100	162
100	500	100	81

The complexity of different Computational tasks is presented in the Table 1. Bearing in mind, that Standard Cargo needs at least 1 s to complete we can see that solving a typical task (50 nodes, 50 Ants, 1000 iterations) requires around half a minute to complete on a standard hardware. This makes it evident how much a parallel implementation of the ACO is needed.

3 Related Work on Parallel Implementations of ACO

The long time needed to complete optimization process has prompted researches to propose different parallelism mechanisms just few years after the introduction of the ACO concept [7]. The ants of ACO work on their own, without paying attention to the operation of other ants. However, all of them use the same pheromone array which is constantly updated after the selection of each node by any ant and once more at the completion of an iteration. The key issue is whether the ants use one or more pheromone arrays. The one array solution is close to the original concept of ACO but it requires an intensive communication. To implement it specialized hardware such as supercomputers, clusters of workstations, recently graphics, multicore processors are used. The parallelization with many pheromone arrays require far less communication and hence a greater variety of hardware could use applied e.g. grid environments or even regular computer equipment [8]. A fairly recent taxonomy divides all approaches to parallel implementation of the ACO is given in [9].

The proposed approaches could be divided into three broad groups:

- Master-slave modes,
- Cooperative models and
- Hybrid models.

The basic features of the taxonomy are summarized in the Table 2. The D/P abbreviation stands for depending on proposal.

Table 2 Characteristics of the models in the new taxonomy [9]

Model	Population organization	# Colonies	# Pheromone matrices	Communication frequency
Coarse-grain master-slave	Hierarchical, non-cooperative	One	One	Medium
Medium-grain master-slave	Hierarchical, non-cooperative	One	One	Medium-High
Fine-grain master-slave	Hierarchical	One	One	High
Cellular	Structured, cooperative	One	Many	Medium
Parallel independent runs	Distributed, non-cooperative	Several	Several	Zero
Multicolony	Distributed, cooperative	Several	Several	Low
Hybrids	Hierarchical	D/P	D/P	D/P

4 Ant Colony Community

The Ant Colony Community (ACC) belongs to the class of hybrid models. It consists of several colonies like the multi-colony models but additionally it has also is a separate server that coordinates, distributes and integrates tasks executed by individual colonies. The ACC was presented before in [6]. The version described in this paper extends considerably the initial concept. The colonies interact now in a more elaborate way and changes in the structure of the community at runtime are also possible. On the implementation level the communication the Sockets has replaced more flexible but slower Remote Method Invocation (RMI).

4.1 ACC Structure

The ACC consist of a single server and a set of colony clients. The communication uses socket mechanism so the ACC components could run a single computer, many computers in a local network, on internet servers or any combination of the above locations. The Colony Client is executed as a separate process and has its own copy of JVM. The client handles all communication with its server. The Client is responsible for creating an object that implements the basic version of an Ant Colony Optimization algorithm [10], passing the input data and operational parameters to it and finally receiving results and sending them to the Server. The community Server has a separate thread for each of the colony clients. It is responsible for allocating tasks to colonies and for integrating the results sent by Colony Clients. An exemplary structure with 4 computers and 5 Ant Colonies is presented on the Fig. 1.

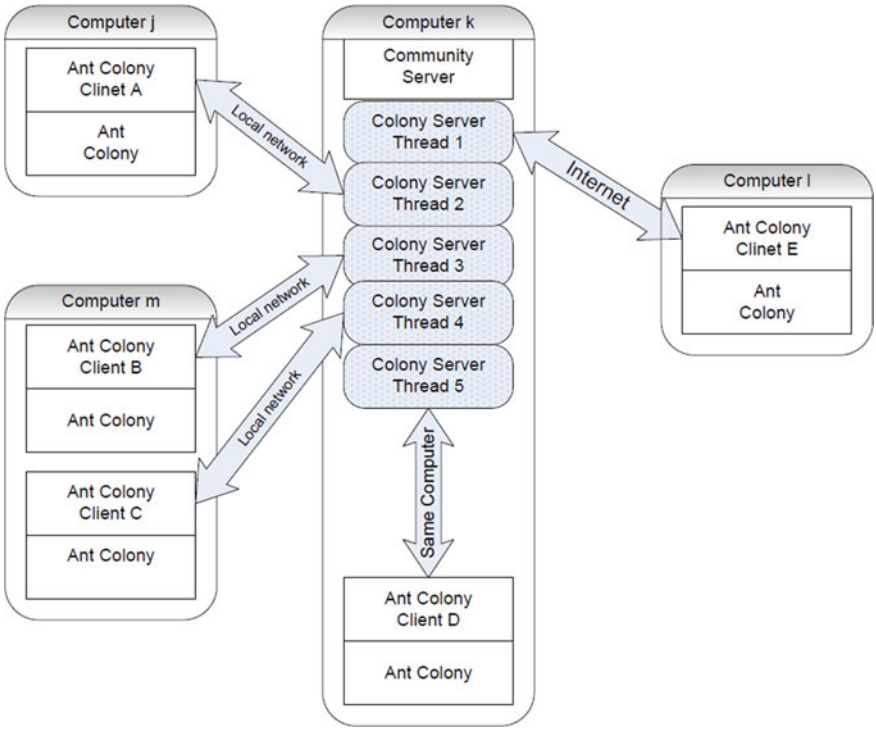


Fig. 1 An example of an ant colony community [6]

4.2 ACC Operation

The interaction between a Server and a Cargo has to be initialized by a Client since the IP address of the server is a property of the community. The server creates a thread dedicated to the calling in client and registers it. After the initialization phase the Server and a client exchange data items which are serialized objects of the *Ant Community Cargo* type. The object contains all data necessary to perform optimization process. The Server starts the work of the Community only after a predefined number of clients have registered. The Colony Client after receiving an *Ant Community Cargo* object creates a local ant colony, feeds it with the data extracted from the just received cargo object and finally starts it operation. An Ant Colony produces a solution in a usual manner and passes the obtained results to Ant Client which packs them into a cargo object that is transmitted further to the Community server.

The rules for message passing between the server and a client are described in the Table 3. The server operates in one of the two following modes:

Table 3 The message passing between a server and a client colony

Community server operation	Ant colony client operation	Remarks
Stop if work is accomplished		The predefined number of objects has been processes by clients
	← Register Client	A separate thread is created to handle the Client. Register data includes: Client Identifier and location, current time. They are stored in a server's registry
Sending cargo to a client →	**Using data from server the client creates an Ant Colony and starts its operation**	**Cargo with raw data to process**
Result integration	← **Found solution**	**Cargo with results of optimization**

- Monitor mode in which the pheromones and partial solutions are not propagated and as a result the clients work independently. The Server limits its activity to monitoring the obtained results and selecting the best one.
- Co-operative mode in which it keeps the partial results in a store. The size of the store is fixed during the optimization run.

5 Efficiency Measures of ACC

In this section we describe the level measures of community performance. Low level measures power and scalability describe the capability of processing data.

5.1 Power and Scalability

The two measures are based on complexity of the task that is worked on. The *Cp* (Computational Power) measure for the Community *Com* and the task *T* it is defined by the following ratio:

$$Cp(Com, T) = \frac{Sc\,(T)*60}{Time\,(T)} \tag{2}$$

where:

Com—the tested community
$Sc(T)$—the complexity of optimization task *T* measured in CP units.
$Time\,(T)$—the time necessary to complete the task *T* measured in seconds.

In other words the Computational Power specifies the number of Standard Cargo tasks that could be calculated within a minute.

The power of the community comes from individual clients. The maximal power that a client *CL* can contribute to its colony is approximated by the $MaxPowr(CL)$ function. It is the power of a colony constructed in the following way:

(a) *CL* is the only one Client Colony in a Community.
(b) The server is located on a LAN computer.
(c) The server does not run any computing intensive tasks.

Experiments have shown that neither the nature of the tasks run on *CL* nor the properties of the server have much impact on the resulting power. Nevertheless all its values shown in the Table 4 were calculated while processing standard cargos. The power measure depend only on time and task complexity so it reflects the actual computational capabilities of the computer hosting a client colony.

It is possible to run several instances of a client colony on a single computer. During the experiment a number of different computers were tested. They were regular desktops, laptops and a university workstation. In order to be able to configure reasonably the structure of a Community we have to know the power of colonies of with increasing number of client colonies. Each colony runs using a separate JVM what drains the recourse pretty fast but still for a small number of client an increase in power is evident, see Table 4.

As you can see from all of the computers were equipped with multicore processors and several GB of memory but their power differs considerably in extreme case by the factor of 3.6. Moreover the difference grows with the increasing number of hosted colonies. In all cases running more than 4 clients hardly increases the community computational power. The university work station is a special case as it hosted many virtual environments whereas all other computers were dedicated to the experiment.

Having defined the *MaxPower* we can measure the $ScalFactor(Com)$ scalability factor for the community Com. It is the ratio of the actual measured power while solving a task and the *MaxPower* sum of its clients.

$$ScalFactor(Com, T) = \frac{Cp(com, T)}{\sum_{Cl \in Com} MaxPw(Cx)} \tag{5}$$

The scalability of tested computers is shown in the Table 5.

Table 4 The Computational Power for communities with differing number of clients, LAN server

Computer type	Number of client colonies				
	1	2	3	4	5
Laptop 1	16.03	29.46	32.44	32.44	32.26
Laptop 2	51.62	98.50	132.40	159.50	164.30
Workstation	39.52	74.41	84.91	99.47	99.48
Desktop 1	16.41	32.82	44.70	62.88	72.12
Laptop 3	57.73	102.63	127.63	146.08	167.84
Desktop 2	21.53	43.37	54.83	62.47	68.78

Table 5 The ScalFactor for communities with differing number of clients, LAN server

	Number of client colonies			
Computer code	2	3	4	5
Laptop 1	0.92	0.67	0.51	0.40
Laptop 2	0.95	0.85	0.77	0.64
Workstation	0.94	0.72	0.63	0.50
Desktop 1	1.00	0.91	0.96	0.88
Laptop 3	0.89	0.74	0.63	0.58
Desktop 2	1.01	0.85	0.73	0.64

6 Configuring ACC

The data shown in the Table 5 refer to communities in which all clients share the same computer. This is rather unusual but the data could be used to help us optimize the setting up communities using many computers.

There are two factors that should be taken care of:

- Computational Power: the large it is, the faster results are obtained.
- Scalability: large values of it indicate that the computers hosting clients are still capable of running some other tasks.

Finding a proper community requires setting acceptance criteria and checking all possible communities. In what follows the term community means an actual instance of ACC whereas the term configuration is reserved for a theoretical community defined by named set of clients. The performance of a community is measured whereas that of configuration is calculated. Using even a few computers the number of all possible configurations could well exceed several hundred. Therefore an algorithm is needed to evaluate all of them.

The properties of computers are represented by two dimensional arrays power and scale. The number of rows and columns in each row must be the same in both arrays. The number of columns could differ as in the below example:

```
double [][] power= {{10.5}, {8.0, 12.0}, {4.0}};
double [][] scale={{1.0}, {1.0, 0.66}, {1.0}};
```

Providing only one value for the first row means that we want to limit the number of client colonies running on the first computer to 1 so it would not be overloaded by the work on behalf of Community.

The number of all possible configurations is given by the method *number of Configurations*:

```
public int numberOfConfigurations() {
        int result=1;
        for (int comp=0; comp<power.length; comp++) {
                result*=(power[comp].length+1);                }
        return(result);          }
```

The key method maps an integer in the range from $[0, \ldots, number\ of\ Configurations() - 1]$. into an integer array that specifies the number of clients hosted on consecutive computers:

```
public int [] config(int conNo) {
        int res[]= new int[power.length]; int reminder;
        int quotient=conNo;
        for (int k=0; k<res.length; k++) {
                reminder=quotient % (power[k].length+1);
                res[k]=reminder;
                quotient=quotient/(power[k].length+1);
        } return(res);  }
```

As you can see the algorithm implementing the mapping is very much like the standard algorithm for changing the base of a number system of an integer value representation. The important difference is, that the base is not fixed, it depends on the length of element in the rows. The calculation of a configuration power and scalability is straightforward. The 4 configurations with the maximum power for the exemplary data are shown in the Table 6.

The configurations 2 and 3 have the same number of clients and power but in in the latter one disperses computations more evenly across the its pool and as a result it has better scalability. The last one has the highest power that comes at the cost of possible overloading of the computers.

In order to verify the usefulness of the configuration analysis an experiment was conducted. It used 4 computers. One of them hosted only the Community Server which was available to the clients via LAN. The arrays describing the power and scalability of the remaining 3 computers are given below.

The total number of all configurations is 216 and maximal configuration power is 402.17. The Table 7 compares the computer power of selected configurations with the actual power of the Communities that they describe.

Table 6 The power and scalability of selected configurations

Number	Power	Scalability	Computer workload
1	18.5	2.00	[1 1 0]
2	22.5	1.66	[1 2 0]
3	22.5	3.00	[1 1 1]
4	26.5	2.66	[1 2 1]

Table 7 Comparing predicted and actual power of communities

Configuration			ACC
Power	Scalability	Computer layout	Power
299.80	0.181	[0 4 4]	275.18
274.35	0.381	[1 4 2]	249.17
248.80	0.332	[4 2 2]	227.39
241.64	0.455	[1 3 2]	219.67
182.51	0.57	[2 2 1]	176.99
159.13	0.715	[1 1 2]	157.87

7 Conclusions

Ant Colony Communities offer an efficient way of parallelizing the ACO applications. Parallel work of many Colonies make it possible to mitigate two main disadvantages of the ACO implementations: long processing time and diversity of results. The version of ACC described in this paper enables us to distribute the optimization job over a large number of individual ant colonies located in nodes spread on a LAN or WAN network.

The paper introduces two low level measures of the performance the ACC: processing power and scalability. They are used to optimize the structure of such a community. The experiments reported in the paper confirm their usefulness. Using the same computational complexity as a traditional version of the TSP the and ACC could achieve results roughly on the same level. The superiority of the ACC over the traditional approach is clearly visible when the equal complexity criterion is replaced by equal time criterion.

The work on the area of ACC continues and it encompasses: going from static to dynamic TSP [11] and implementing the ACC concept using the Hadoop environment.

References

1. Antosiewicz, M., Koloch, G., Kamiński, B.: Choice of best possible metaheuristic algorithm for the travelling salesman problem with limited computational time: quality, uncertainty and speed. J. Theoret. Appl. Comput. Sci. **7**(1), 46–55 (2013)
2. Dorigo, M.: Optimization, learning and natural algorithms. Ph.D. thesis, Politecnico di Mila-no, Italie (1992)
3. Dorigo, M., Stuetzle, T.: Ant colony optimization: overview and recent advances, IRIDIA—Technical Report Series. Technical Report No. TR/IRIDIA/2009-013 (2009)
4. Wei, X.: Parameters analysis for basic ant colony optimization algorithm in TSP. Int. J. u-and e-Serv. Sci. Technol. **7**(4), 159–170 (2014)
5. Siemiński, A.: Ant colony optimization parameter evaluation. In: Multimedia and Internet Systems: Theory and Practice, vol. 183 5, pp. 143–153. Advances in Intelligent Systems and Computing (2013). ISSN:2194-5357
6. Sieminski, A.: Using hyper populated ant colonies for solving the TSP. Vietnam J. Comput. Sci. **3**, 103–117 (2016)
7. Randall, M., Lewis, A.: A parallel implementation of ant colony optimization. J. Parallel Distrib Comput. Academic Press Inc **62**, 1421–1432 (2002)
8. Delévacq, A., Delisle, P., Gravel, M., Michaël Krajecki, M.: Parallel ant colony optimization on graphics processing units. J. Parallel Distrib. Comput. **73**, 52–61 (2013)
9. Pedemonte, M., Nesmachnow, S.: CancelaH.: a survey on parallel ant colony optimization. Appl. Soft Comput. **11**, 5181–5197 (2011)
10. Chirico, U.: A Java framework for ant colony systems. In: Ants 2004: Forth International Workshop on Ant Colony Optimization and Swarm Intelligence, Brussels (2004)
11. Soleimanian, F.: New approach for solving dynamic travelling salesman problem with hybrid genetic algorithms and ant colony optimization. Int. J. Comput. Appl. (0975–8887) **53**(1) (2012)

Detection of Security Incidents in a Context of Unwelcome or Dangerous Activity of Web Robots

Marcin Jerzy Orzeł and Grzegorz Kołaczek

Abstract This work presents several scenarios used to identify security incidents based on the analysis of web server log files. The main goal of this work is to identify security events triggered by web robots which can be considered as dangerous or unwelcome. Analysis of all security incidents was based on archived web server log files which were collected from 03.03.2014 to 31.01.2015 and came from the real and fully functional environment, available at www.darmowe-obrazki.pl. All data were obtained automatically on a daily basis and analyzed using Advanced Web Statistics software.

Keywords Web robots · Web crawlers · Web security · Privacy · Incidents

1 Introduction

Web robots also known as Web Wanderers, Crawlers, or Spiders have been used in Internet over two decades. It is assumed that appearance of the first web robot happened in 1993. It was the World Wide Web Wanderer and briefly commonly also called as the Wanderer. The main goal of its activity was to measure the growth of the Internet and to extract a number of active HTTP pages [5]. In general, the main tasks of robots can include activities such as collecting data for the search engines, validating code of the web pages, collecting information and monitoring changes of the web site as well as the creation of web sites mirrors, it means providing the exact copies of indexed web pages in a different location of Internet.

M.J. Orzeł (✉) · G. Kołaczek (✉)
Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wrocław, Poland
e-mail: Marcin.Orzel@pwr.edu.pl

G. Kołaczek
e-mail: Grzegorz.Kolaczek@pwr.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_19

The most important thing which should also be noticed that the details of the operation performed by web robots are not published. In many other cases the available information of the web robots is outdated or inaccurate.

While Internet has become more and more popular and also more and more data was available in web pages the range of the application for web robots increased. Together with the increased web robot activity, issues related to privacy are becoming more visible. The natural consequence of this was an urgent need to develop a system which would allow getting control over the actions performed by web robots. In 1994, the Dutch developer Martijn Koster [7] created the so-called "Robot Exclusion Standard" [3, 4], which laid down the principles of access to server resources through the commands placed in the file named robots.txt [3]. This de facto standard is still used as a basic mechanism for control the activity of web robots. Next attempts to normalize the activity of robots were taken by the W3C consortium which published in December 1999 the HTML 4.01 specification containing guidelines regarding the use of META tags [1].

High activity of robots can also be a source of problems for web server administrators as well as common users publishing in the Internet. At least this type of activity could be viewed as an unrequired. The smallest problem is, of course, increased usage of server resources, especially in cases while the web pages contain a lot of graphics or multimedia. But some web robots may also perform actions which can bring some harmful effects on web server. For example, web robots may be used to create so called botnets to carry out Internet DDoS attacks, collect private or secret data by phishing, perform clickjacking attacks, etc. Currently, the problem of "bad" use of robots also applies to the world of entertainment (games) and is referred to as a farming. Farming can be performed both by humans and by a specially designed for this purpose software. The phenomenon involves repetition over a period of time some types of activities, in order to achieve significant benefits for example in the virtual world of the game. This type of activity is generally not accepted because it disturbs the equilibrium between players.

The goal of the paper is to present the authors' methods of testing the behavior of web robots and discussion over a proposed taxonomy of web robots where robots are classified using traces of their activity in web server. In particular, studies have focused on identification of information security problems associated with the activity of web robots.

The next section contains the proposal of five unique scenarios defining a way to monitor the activity of the web robots in the context of the predefined characteristic elements. Selected scenarios are described in detail and are illustrated with examples of the results coming from observations of real web traffic and web robot activity in a research environment. The last part of the paper contains a summary of the work and proposals for further actions related to the issues of security and web robots.

2 Taxonomy of Web Robot Malicious Activity

Network traffic related to the www.darmowe-obrazki.pl website, was monitored from 03.03.2014 to 31.01.2015 on a daily basis using CRON jobs and Advanced Web Statistics software (see Fig. 4). During this time, exactly 600 log files were created with the total volume of 20 MB. Using the experience form the observed characteristics of the traffic generated by web robots ten unique test scenarios have been proposed. These scenarios can be used for the detection of security incidents caused by, among others, undesirable activity of web robots. Scenarios presented below have been implemented and verified in the research environment which was constituted by a typical web server connected to the Internet. However, proposed method is universal and may be applied to another environment, e.g. to the software development environment or used during next research stage on the web robots activity.

- S01—Server error No. 500
- S02—Reading the robots.txt file
- S03—Compliance with META robots tag
- S04—Direct request of pages referred at homepage
- S05—Excessive consumption of server resources by web robots.

2.1 S01—Server Error No. 500

The first scenario describes the effects of an HTTP error No. 500 caused by excessively frequent requests to the HTTP server. The term "too many request" can be defined as the situation when at least several dozen requests per second are sent to the server. In the test environment, the maximum recorded number of requests made within 1 s by a web robot amounted to 43.

The study has shown that frequently sending requests to the server by web robots can lead to an error No. 500 (Internal Server Error). Using this type of error an attacker can learn some secret data e.g. private e-mail of service provider or owner of the resources. The occurrence of this error can also be related to some server misconfiguration. In the test environment error no. 500 was observed two times. The occurrence of these errors was caused by sending food requests to the server by a robot Nikto.

Another problem related to this scenario is a way how to detect this error. As this error is defined as "unexpected problems which prevented completion of the request" then the request that caused this error may not be registered in the Web server log file. For making detection of this error more fast and accurate, we can use the statistics system like e.g. AWStats, which offers monitoring of HTTP error codes. A good practice is also to use as the default e-mail address for server

administrator, the address not unveiling any personal data. For example the e-mail address should be better neutral as administrator@example@com or webmaster@example.com, than first_name.last_name@example.com.

2.2 S02—Reading the Robots.txt File

The second scenario describes a security incident when web robot starts a web page indexation before reading the robots.txt file. The first step of each web robot should be to download the robots.txt file in order to read the directives which can allow or forbid indexing some specific part of a Web portal content. This operation is obligatory that web robots should absolutely do before beginning their main activity [3, 4]. The administrator/owner of the website does not have power to verify whether the guidelines from the file robots.txt are respected. However, using the file server logs files administrator can determine whether a specific user coming from some specific IP address made a request for downloading the robots.txt file. This is the basic operation which allows for the classification of robot in terms if it follows a set of web systems standards. Web robots that do not load and do not follow the guidelines included in robots.txt certainly can be classified as undesirable because they lose control over their activity and it is not possible to have the effect of reducing the scope of their activity through the basic mechanisms used by web servers. E.g. in such cases it is not possible to permit retrieval and indexing of specific resources of website using standard *Disallow* directive.

The following records from event file (Fig. 1.), illustrate the process of accessing the robots.txt file performed by Googlebot before it started the website indexation. This is an example of good behavior of web robots.

This is an example of proper behavior, because the request to the robots.txt file was made first. In the situation when robots.txt included a directive providing for disagreement on indexation of the website content, web robot should immediately stop sending subsequent requests to the server. Web robot may then access only

```
66.249.78.152 - - [13/Oct/2014:03:04:47 +0200] "GET /robots.txt
HTTP/1.1" 200 274 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"

66.249.78.152 - - [13/Oct/2014:03:04:47 +0200] "GET / HTTP/1.1"
200 1487 "-" "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS
X) AppleWebKit/536.26 (KHTML, like Gecko) Version/6.0 Mo-
bile/10A5376e Safari/8536.25 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html
```

Fig. 1 Googlebot activity log

robots.txt file in order to verify whether such a directive has not been deleted or changed.

For robots that do not read a robots.txt file and do not follow the instructions contained therein, an effective way of defense is to block address or IP address pool from which the robot comes from. When the robot uses constant user-agent metric, it is possible to use some software to filter out unwanted traffic or redirect it to some predefined resources which may be a type of honeypot. Unfortunately, this information used by web robot can be changed as easily as a different IP address can be used [9]. For the best results combining several techniques to filter network traffic can be used. For example, taking into account a few different criteria for identifying unwanted web robot. Another approach may be use of CAPTCHA or other user authentication mechanisms.

2.3 S03—Compliance with META Robots Tag

The third scenario describes the incidents involving the failure to respect the robots META tags, such as *noindex* and *nofollow* (Fig. 2). With the correct use of META tags it is possible to define some rules determining behavior of web robots. Parameter *noindex* is responsible for informing the robots of disagreement for indexation of website content, while the *nofollow* parameter should prevent a processing, viewing, and taking any other action in the context of enumerated elements of website which can be found by web robot while parsing the web page.

The above-mentioned parameters should be considered rather as recommendations for web robots and not as something what is guaranteed and what can determine and limit the actions undertaken by web robots. This is because the part which is responsible for respecting the META robots is the creators of web robots. Finally only implemented within web robot mechanisms for handling these tags define the way how these tags are interpreted. A good example of the different interpretation of *nofollow* parameter can be noticed in the context of leading search engines. The differences start at the interpretation of the meaning of the term *nofollow*. Currently, there is a tendency to try to index each element of a website and then to remove or mask forbidden content which was identified as *noindex*/*nofollow*, so the results returned by a search engine do not included prohibited elements.

```
<head>
        <title>Darmowe obrazki</title>
        <meta name="description" content="Super kolekcja darmow
        <meta name="keywords" content="darmowe, obrazki, przyrc
        <meta name="robots" content="noindex, nofollow" />
</head>
```

Fig. 2 Noindex and nofollow example

Detection of such incidents is difficult and usually takes place a posteriori. During the study it was found that a very promising method is to use a webpage footer which after indexation will appear in the online copy made by a web robot. This gave the opportunity to compare the time for the operation, configuration of the environment with the audited period. The condition is to create and maintain the repository of all versions of the website code. In addition, the proposed method allows avoiding dependence on the information provided by the search engine provider. Please note that the webpage footer must be visible in the page content. The hidden footer should not be used in this purpose (e.g. as a comment in the code page) because the footer of this type may be omitted in the process of creating a webpage copy. In order to detect the real-time scanning of a page marked as *nofollow* can be used a hyperlink to a dedicated and monitored resource for example honeypot.

2.4 S04—Direct Request of Pages Referred at Homepage

The fourth scenario describes the direct requests to pages linked at homepage. During the analysis of the Apache web server log, the following situation has been identified in the test environment. Sometimes selected files which were used to compose the main webpage "www.darmowe-obrazki.pl" were uploaded by the web robot (Fig. 3). The layout of the main webpage suggested that during each web

Fig. 3 The main webpage of the test environment www.darmowe-obrazki.pl

robot visit at least nine pictures should be uploaded by the robot. In practice the observed behavior of the web robots was different.

The main web page is generated u sing a pseudorandom mechanism. During each visit another set of nine images is chosen and displayed. It was confirmed that in the case of addition new pictures to the web page, all pictures linked at the main web page will be downloaded by web robots. However, during next visits, web robots usually take only selected files, often without the request to the root of the domain, or to index.html or index.php files.

For several years, there are theories [2] where it is assumed that web robots can actually be browsers that have been suitably modified and adapted to automatically download the content. These theories are complimentary with the observed web robot activity described earlier. It means that the web robot may have some kind of cache memory—similar to cache memory implemented in the web browsers, which is used by web robot exactly as it is used in web browsers installed on personal computers.

It was observed that the web robot could download the same files repeatedly. For example some jpg files which were not modified, were downloaded on the same day in the morning and again in the evening. Due to non-disclosure policy which is common for web robots developers it can be assumed that this behavior may be related to task of indexing only new elements which is so-called "fresh crawl" or indexing all elements, which is referred as "deep crawl". There is also possible relation between network traffic generated by web robots and the positioning of the resources. However, one should also be aware of the negative aspects of extensive web robot activity. Generating too large network traffic could result in blocking of the selected web page. That vulnerability relates primarily to web pages that offer rich multimedia content like high-resolution images or video clips.

While the available monthly data transfer rate has been set at 10 GB, the largest monthly traffic generated by a web robot was registered in November 2014 and amounted to nearly 70 MB (69.23 MB) (Fig. 4). In the case of the testbed which was used during our research, single download of a portal home page "www.

Roboty sieciowe

14 różne roboty sieciowe	Żądania	Pasmo	Ostatnia wizyta
Googlebot	368+37	69.23 MB	30 Lis 2014 - 23:48
Yandex bot	85+39	6.75 MB	30 Lis 2014 - 21:58
Unknown robot (identified by 'bot/' or 'bot-')	63+16	19.53 MB	30 Lis 2014 - 15:10
BSpider	49	3.39 MB	30 Lis 2014 - 23:48
MSNBot-media	20+18	8.36 MB	13 Lis 2014 - 03:35
MSNBot	9+9	3.27 MB	12 Lis 2014 - 08:26
W3C Validator	14	31.80 KB	30 Lis 2014 - 16:27
Unknown robot (identified by hit on 'robots.txt')	0+12	3.08 KB	30 Lis 2014 - 16:44
BaiDuSpider	11	25.04 KB	27 Lis 2014 - 00:11
Python-urllib	5	11.19 KB	30 Lis 2014 - 15:48
SurveyBot	2+2	5.06 KB	24 Lis 2014 - 19:01
archive.org bot	1+1	2.54 KB	06 Lis 2014 - 11:03
W3C jigsaw CSS Validator	2	4.62 KB	30 Lis 2014 - 15:35
ParaSite	1	2.25 KB	30 Lis 2014 - 16:05

* Pokazane tutaj roboty sieciowe generują ruch "nieoglądany" przez gości, dlatego nie są zawarte w innych zestawieniach. Liczby po + dotyczą udanych odwołań do plików "robots.txt".

Advanced Web Statistics 6.7 (build 1.892) - Wygenerowane przez awstats

Fig. 4 The level of data transfer observed in the test environment on November 2014

darmowe-obrazki.pl" generated data transfer of ~3.5 MB. It means, that to block portal, only about 2900 refreshes of the main page per month would be enough. And this is equal to less than 100 hits on this web page throughout the day. The calculation very clearly demonstrates the vulnerability of low-budget hosting platforms to attack aimed to exhaust server resources. This security incident may be linked also with another phenomenon called hotlinking.

Hotlinking is based on the use of resources on a different server than the site visited. A good example is a blog that does not have its own pictures, but downloads them from another server, e.g. from "www.darmowe-obrazki.pl". This situation also can be noticed during web server logs analysis. It will be denoted by logs where selected files from our web page are downloaded without corresponding requests to the home page of the web site. Hotlinking for resources may be useful as a form of advertisement. But it also may be treated as unwanted activity if it is used without the agreement of the owner of the site that contains the downloaded files. This is a serious incident that can be classified as stealing bandwidth. To protect against this threat, special programs can be used to analyze the HTTP header in order to determine whether a request for downloading the resource came from a party involved in the same domain as the location of the requested resource.

2.5 *S05—Excessive Consumption of Server Resources by Web Robots*

The scenario S05 considers the problem of excessive consumption of server resources by web robots. A few not standard rules were included in robots.txt file to force robots to return to the research environment at certain interval of time. Also the technique of automatic forwarding using META refresh tag has been analyzed. The problem arises when robots scan the site excessively and when scans are repeated frequently in short period of time. The observations of the test environment have shown that some web robots are extremely active and use of server resources more than the others. This observation motivated the search for the mechanisms which would allow reducing such a type of behavior. In the case of continuous indexing the unchanged content of web, the website owner cannot expect benefits from web robots activity. To the contrary, the site owner may lose. For example in the case of exhaustion of the of data transfer limit, a service of regular users can be blocked and owner of the site is charged due to necessity of buying a larger data transfer limit. Apart from the standard commands that control the work of robots [1, 3, 4], there are also some new commands, extending control over the web robots activity. The first novelty which has been introduced is the parameter Crawl-delay in order to establish in seconds the time interval that should pass before the same web robot again sends the request to the server in order to perform indexing of the site content (Fig. 5).

```
User-agent: *
Crawl-delay: 86400
```

Fig. 5 Example of directive Crawl-delay with delay set to 24 h

```
User-agent: *
Allow: /obrazki-dla-kazdego
```

Fig. 6 Example of directive Allow

Another new element is Allow directive. The purpose of this directive is explicit indication of resources that should be indexed by web robots (Fig. 6).

The new directives are not recognized by all web crawlers, but at least Googlebot and Bingbot support this approach [8]. In order to maintain compatibility with the interpretation of the directives, it is recommended to put directives Allow before directives Disallow. The example shown in Fig. 7 allows defines the situation where only one file of the entire directory should be accessible to web robots.

Another way to reduce the excessive activity of web robots, may be a method involving the automatic generation of headers, causing forwarding unwanted traffic to a separate part of the server (honeypot like or web page without multimedia content). The idea is not to block scanning, but reducing the amount of data transferred. Because scanning of the web site influences the position of the site in web page ranking it is not recommended to totally block the web robots. However, sometimes it may be recommended to limit their activity (for example in rush hours) using *META refresh* directive (Fig. 8).

The proposed method, although simple will effectively influence the activity also on the web robots that do not respect the directives contained in robots.txt file. Initiating factor redirecting traffic can be for example: the name of a web robot, user-agent metric, IP address, reputation of the robot, etc. Modified headers can be created with a script written in PHP page but there are also more advanced methods as for example based on the settings of the Apache server (.htaccess) using the mechanism of mod_rewrite, as well as using HTML frames or CGI programs written in Perl.The performed experiments in the test environment using several

```
User-agent: *
Allow: /obrazki-zastrze.one/obrazek-001.jpg
Disallow: /obrazki-zastrze.one
```

Fig. 7 Example of Allow and Disallow directives in single robots.txt file

```
<meta http-equiv="refresh" content="0;
 url=http://www.darmowe-obrazki.pl/honeypot-dla-robotow">
```

Fig. 8 Example of META refresh directive

different values of the above mentioned directives resulted in the following observations. The modifications of the directives and blocking the access to the web site for robots bring a web page address removal from the search engines results. Another interesting result was the case of a web robot MJ12bot where Crawl-delay value can be at most 20 s [6].

3 Conclusions and Future Works

During the research and literature overview, it turned out that all available materials describing the characteristic of web robot activity are incomplete or outdated. It should also be noticed that despite the availability of basic information about web robots including such as the name of the owner/developer of the robot, the information on their specifications and rules of operation are in most cases not available. The presented results in this paper might be used to detect UNWELCOME or DANGEROUS activity of web robots with particular interest on the risk situations, violations of data security during the actions performed by robots while collecting information about structure and resources of web servers. The classification of the robot's activity can be done using only its activity traces so even if the exact name of the robot is unknown. The future works will focus on application of non-standard directives such "allow" or "crawl-delay" e.g. for creation of dedicated honeypot and for design of "Model for web robot classification". This will be done using web robots activity patterns described in this paper. As the test environment is planned to be available in the Internet till the end of 2016, it is possible that some new problems and emerging trends in web robots development will be detected which will be described in details in next publications.

References

1. HTML 4.01 Specification—Appendix B: Performance, Implementation, and Design Notes—B.4.1 Search robots, W3C (1999). http://www.w3.org/TR/html4/appendix/notes.html#h-B.4.1.1
2. Josh, U.A.: Googlebot is Chrome (2011). http://ipullrank.com/googlebot-is-chrome
3. Koster, M.: A Method for Web Robots Control, Network Working Group, Internet draft (1996). http://www.robotstxt.org/norobots-rfc.txt
4. Koster, M.: A Standard for Robot Exclusion, Internet draft (1994). http://www.robotstxt.org/orig.html
5. LaMacchia, B.A.: Internet fish. Ph.D. thesis, Artificial Intelligence Laboratory and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (1996). http://www.farcaster.com/papers/ifish/ifish-tr.pdf
6. Majestic-12: DSearch: MJ12bot—How can I block MJ12bot? (2014). http://www.majestic12.co.uk/projects/dsearch/mj12bot.php
7. Martijn Koster—Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Martijn_Koster. Accessed July 2014

8. Robots exclusion standard—Nonstandard extensions—Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Robots_exclusion_standard#Crawl-delay_directive. Accessed Feb 2015
9. Scrapy 0.24.4 Documentation—Settings (ROBOTSTXT_OBEY). http://doc.scrapy.org/en/latest/topics/settings.html. Accessed Jan 2015

Stereoscopic 3D Graph Visualization for Assisted Data Exploration and Discovery

Michal Turek, Dariusz Pałka and Marek Zachara

Abstract Data structures and relations are becoming increasingly complex and difficult to assess and manage. Although automated rules and algorithms can be used for many data-mining tasks, there are still situations where human attention and insight is required to identify unexpected circumstances or unanticipated patterns. Presentation of large quantities of data has always been a challenging task. In this paper a method for representing large graph-based data sets is proposed to help users navigate through large clusters of data. The proposed method is based on a stereoscopic 3D visualization with special enhancements for a large multi node graph visualization. The stereoscopic projection allows for utilization of techniques that can draw users' attention to particular regions of the graph. The method uses specially established node-node relations to calculate attention drawing factor values for each graph node.

Keywords Stereoscopy · Visualization · Graphs · Data exploration

1 Introduction

Exploration of large data sets is not an easy task. While there have been numerous methods invented for identification of patterns and data mining in such data sets, automated methods have problems with identifying unanticipated patterns or irregularities. This is the area where human skills are often superior and still required. On the other hand, humans cannot usually analyze or operate on large data sets directly, and they require the data to be pre-processed, e.g. aggregated, clustered and/or

M. Turek · D. Pałka · M. Zachara (✉)
AGH University of Science and Technology, Krakow, Poland
e-mail: mzachara@agh.edu.pl

D. Pałka
e-mail: dpalka@agh.edu.pl

M. Turek
e-mail: mitu@agh.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_20

classified before it can be presented to the operator. This paper explores the benefits of utilizing stereography and 3D rendering for presentation of large graph-base data sets. As a proof of the concept, the method was used to analyze and identify irregularities in a large web-site (a news portal) structure. The proposed method can be used both in real-time rendering of animated 3D presentations as well as in applications where static stereoscopic pictures are needed. In both cases there are two projection windows used for both eyes of the observer. Each image is generated with a 4×4 projection matrix [6], with images displaced relative to each other. This results in a stereoscopic effect and is a standard and well known method [17]. In a proposed solution both views are shifted dynamically multiple times during rendering of each animation frame.

2 Related Work

Node-link based graph 3D visualization methods are widely deployed in many tools and math packages e.g. Statistica or Mathematica. However, visualization techniques used there rely on a single-image 3D graphical presentation. Stereoscopic graph visualization techniques with advanced node processing is still a significantly undeveloped field. Attempts have been made to process plain 2D images for stereoscopic graph presentation [14], but the idea there was only to shift important portions (nodes) of a 2D image horizontally between two displays to achieve a stereoscopic attention drawing effect. Earlier theoretical descriptions of the techniques in this field [7] described also different projection types for graph nodes presentation.

Some theoretical work has also been done on depth rendering optimization [9]. The methods described allow for tuning of optimal graph node stereoscopic shifts in regard to virtual distance, targeted projection method and high-level factors (importance, activity, etc.). Additionally, some engineering-ready graph rendering methods are available which utilize stereoscopic highlighting of a particular graph's node by projecting it slightly closer to the observer [1]. However, with no stereoscopic depth calculations involved, the bethod utilizes only classical distance-based transformations to render the stereoscopic signal. Diagrams or other node-link structures can also be highlighted in this way [2].

3 Overview of the Method

A 3D position and orientation of both virtual cameras is set to a view somewhere in the projection space [12], determining an optimal stereoscopic projection depth (usually called a stereoscopic projection plane). In other words—a stereoscopic projection plane is placed in the exact distance from the cameras where (in an ideal situation) the presented 3D object should be.

The goal is to provide Sable mathematical formulas that can be used with 3D hardware accelerated libraries such as OpenGL or Direct3D [12]. In order to support such application, 3D transformation matrix formulas will be used to calculate the stereoscopic projection transformation with projection plane modifications implementing the discussed method [5]. In fact, there are two matrix structures needed—one for the left eye projection, and another one for the right.

To begin with, a matrix definition is needed. It is a 4×4 rectangle matrix of float numerical values [13]. The goal is to achieve a native form of the matrix that can be used with a 3D hardware accelerated rendering process. When aligned, any kind of elementary 3D object component position vector (3D point in most cases) will be multiplied by the matrix, resulting in a new position on the screen. That emulates the positioning of the "camera" and any kind of other projection window change. In the proposed method, there are two projection windows and two matrix transformations used for the generation of modified views for both eyes. To create the projection matrix content, the projection cameras layout must be fixed. The cameras are positioned on the X-axis, simulating the relative location of the eyes. They are oriented towards a point on the Z-axis—indicating a standard stereoscopic position plane distance. Considering the stated goal, the most important factor (and also an input parameter for the matrix creation) is the stereoscopic projection shift—applied as a correction of the point position, positioned on the Z-axis at $(0, 0, -pointZ)$ and observed by two cameras. Additionally, a vector pointing camera-top for both cameras must be defined as an input parameter. All those assumptions lead to the following list of 3D input vectors: LC—left camera position, LP—left camera look-at point position, LT—left camera top vector, RC—left camera position, RP—right camera look-at point position, RT—right camera top vector:

$$[L/R]\ C = \begin{bmatrix} -eyeSpread \\ 0 \\ 0 \end{bmatrix} [L/R]\ P = \begin{bmatrix} 0 \\ 0 \\ -(pointZ + shift) \end{bmatrix} [L/R]\ T = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

First, the camera position and the camera look-at point vectors subtraction are calculated as follows:

$$LD = LP - LC, RD = RP - RC \tag{1}$$

Then, both vectors are normalized:

$$LD_{norm} = normalize(LD), RD_{norm} = normalize(RD) \tag{2}$$

During the next step, a cross product of the normalized vectors and the initial vectors is calculated:

$$LD_{cross} = cross(LD_{cross}, LT), RD_{cross} = cross(RD_{cross}, RT) \tag{3}$$

and the result is normalized again:

$$LD_{crossnorm} = normalize(LD_{cross}), RD_{crossnorm} = normalize(RD_{cross}) \quad (4)$$

Finally, new top vectors are calculated as follows:

$$LD_{top} = cross(LD_{crossnorm}, LD_{norm}), RD_{top} = cross(RD_{crossnorm}, RD_{norm}) \quad (5)$$

To build the final projection matrix for each of the two eyes, a $4 * 4$ identity matrix must be generated. The vectors *LDcrossnorm*, *LDtop* and *LDnorm* (or *RDcrossnorm*, *RDtop* and *RDnorm*) must be placed respectively in the beginning of the first, second and third column of the matrix, as shown below:

$$[L/R]\, EyeMatrix = \begin{bmatrix} [L/R]\, D_{crossnorm}(x) & [L/R]\, D_{top}(x) & [L/R]\, D_{norm}(x) & 0 \\ [L/R]\, D_{crossnorm}(y) & [L/R]\, D_{top}(y) & [L/R]\, D_{norm}(y) & 0 \\ [L/R]\, D_{crossnorm}(z) & [L/R]\, D_{top}(z) & [L/R]\, D_{norm}(z) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Constructed formulas are suitable to be applied in the rendering library's viewport positioning measures—resulting in a proper camera positioning for both eyes. They were used as projection matrix transformations for both eyes during the experiment. A "shift" parameter presented above was used to modify the stereoscopic effect strength during stereoscopic perception measurements.

3.1 Experimental Evaluation

The idea was evaluated against a number of volunteers. Each subject was presented with a stereoscopic video signal which had a deliberately miscalculated right-left frame differences caused by a slight discrepancy of the stereoscopic projection plane's distance and a natural 3D object distance. Each person had a choice of several objects projected simultaneously, each with a different projection miscalculation. The person could then mark the one that was most appealing/most correctly looking for him or her. Apart from evaluating static scenes, the subjects were also presented with a similarly constructed motion video, including objects that rapidly increase or decrease their distance to the observer.

Finally, positive (+) or negative (−) stereoscopic focus corrections have been defined as an optimal dynamic shift between mathematically ideal and the modified view actually presented to the observer. Especially good outcome of those enhancements was visible in the case of wire-frame objects rendering, where object parts on the scene are smaller and harder to see, since they are represented only be the contour lines.

Stereoscopic focus is automatically applied for each node or edge during the rendering process. A zero focus value for all the nodes is equivalent to a normal 3D

stereoscopic projection process (no anomalies). As has been discovered, a slight anomaly in node stereoscopic distance rendering gains a human attention. Therefore, some nodes can have an adjusted focus, causing a stereoscopic anomaly. The modification is just a correction of a stereoscopic projection planes distance. Additionally, the focus spreads to another nodes. Each edge has a factor value, expressing two nodes stereoscopic distance and placed between 0 and 1. The next node to be rendered gets the focus from the previous one—multiplied by that distance factor. This happens only if both nodes are connected and have suitable edge existence. Consequently, the stereoscopic focus spreads to another nodes and due to edge factor multiplication—loses its impact. Such approach enables to present focused nodes smoothly and gains observers attention much stronger than a single node anomaly would.

As a result of the described impact of the view corrections on the scene perception by the viewers, the method was considered a prospective candidate for a multi-dimension graph presentation, but it can also be employed to other areas where there is a need to project certain objects from metric space to a 3D Euclidean space. Any kind of a geometrically spread content could be expressed in this way. As one of potential use case scenario, this method has been applied to a website analysis, and the next chapters present an example of using the dynamic projection to identify potential errors or inconsistencies in the analyzed web site.

3.2 Details of the Methodology

As mentioned before, a purposeful change of the objects' mapping from the 3D to binocular (2D) space was used for drawing users' attention to the specific elements of the presented set. This is done on the sub-conscious level, without the need for active user participation. This is an alternative to other attention-drawing methods, such as extreme brightness, specific colour or object movement. In other words, even with all the graph's nodes visually identical, the viewer's attention is drawn to the objects with the modified stereoscopic focus.

Multi-dimension graph visualization tools can process a huge amount of graph nodes, for instance, when storing simulation results within a graph. Those visualizations often focus on the node's attribute value calculations—bounding nodes together and finally expressing a physical force, pollution level, speed etc. bound to each node. It could be either calculated or measured and spread between the nodes, depending on the method. In many cases, 3D graph visualizations show simulation results by expressing those attribute values.

They can also express nodes importance and draw observer's attention. To avoid a situation when the presentation designer has to define the importance level for each node, additional calculations have been added to a proposed rendering method. They are based on a path costing algorithm in a graph. The cost of a path to the nearest node with an importance level set is calculated. That allows for evaluation of node's importance in a real time (during a rendering process). Based on these calculations,

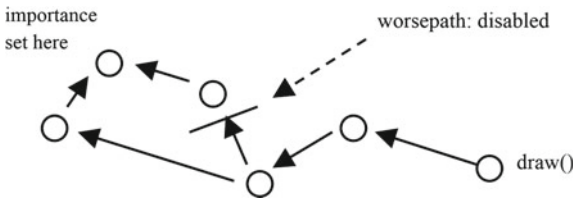


Fig. 1 Graph stereoscopic correction impact calculation

different stereoscopic projection settings are set for each node to be rendered. Subsequently, that *node importance* correction is spread between all the nodes automatically. The graph edge cost (between two multi-dimension graph nodes) can either be set manually or considered as a globally fixed value. It can also be generated as a result of other domain-dependent calculations.

3.3 Node Importance Calculation

In the experimental implementation the node's importance values ware calculated as best path values. Each path cost was calculated as a product of all edge costs on the path that leads from the current node to the node with importance factor value set (Fig. 1).

Each edge cost value has a number in a range between 0 and 1 calculated multiplicatively over the whole path. With this path-based approach, the importance of a certain node is propagated automatically over the whole graph resulting in easier tracking of important nodes by the viewer. There is also an additional shift applied to each node relative to its distance from the designated node in order to let the viewer easily identify the shortest path.

It can express a graph node distance, presented as a natural geometric distance between nodes in the projection space or any other factor values, if needed. When a node importance factor value is calculated, the process simply multiplies a positive (+) or negative (−) stereoscopic focus correction established earlier for the particular person as the best attention drawing value.

4 Case Study

Administration and supervision of a large web site is quite a challenging task. The difficulty is primarily caused by the number of new web pages added, deleted and modified every day. Due to the number of changes, it is quite likely that some web pages that should either be removed or not presented to the public at all are made available to the public. Such incorrect pages may adversely affect the web site usability, the indexing of the web site by search engines, or may even constitute a source of potential vulnerabilities.

4.1 Detecting Anomalies in the Structure of a Web Site

Typical defects found in web sites or web applications are broken links i.e. hyperlinks that point to web pages, services or other resources that are permanently unavailable. These defects can, however, be easily detected automatically: there are numerous programs which detect broken links, e.g. Google Webmaster Tools However, defects that result in exposing incorrect content by a web site are much more difficult to detect. Such incorrect content usually differs considerably from other content of the web site, but cannot be identified with typical web-maser tools described above. There are, however, methods for comparing web pages and identifying the outliers. The method described here relies on a similarity of an underlying HTML code of web pages. During the process of web crawling, all pages of a web site are visited iteratively and the distance (or similarity) between each pair of web pages is calculated. As a result, the set of points in the metric space is obtained.

To calculate the distance (i.e. similarity) between pairs of web pages, the HTML of the web pages are compared using typical text-similarity methods, such as Levenshtein distance [10]. Computing a text similarity like the Levenshtein distance for a raw HTML page code is rather computationally expensive process, the complexity of classical algorithms is $O(n * m)$ [16], and the best known algorithm for constant-size alphabet is only slightly better $O(n * min(1, m/log(n))$ [3]). Therefore, a tokenized version of web pages has been used in the experiment. Such an approach is discussed in [11], and a number of different tokenization methods and text similarity algorithms are compared in [18]. As a result of the tokenization, the size of the compared text is reduced approx. 20–50 times, which greatly reduces the computing time.

Once the distances between the web pages are calculated, the resultant distance graph can be created. This graph is also an equivalent to a metric space.

4.2 Projection from a Metric Space to a 2D or 3D Cartesian Coordinate System

In order to visualize and analyze the structure of a web site or web application so as to find potential anomalies in its structure (which might indicate some deficiencies/defects), the metric space is translated into 2D or 3D Cartesian coordinate system. The transformation used should preserve the distances between objects, i.e. the distance in the Cartesian coordinate system should be equal or possibly very close to the distance between the objects in the metric space. This approach is widely used in Multidimensional scaling (MDS) technique [4, 15].

The spring model, similar to [8], is used in the process of projection of the points from a metric space to 2D or 3D Cartesian coordinate system:

- page hashes (the points in a metric space) are treated as mass points in the Cartesian coordinate system.
- mass points are mutually connected by springs.
- the length of a spring between mass points p_k and p_n is equal to the distance between corresponding web pages w_k and w_n in a metric space. The spring length between points p_k and p_n is denoted $l_{k,n}$.

Restoring force value between a pair of points p_k and p_n is defined in (6).

$$F = -k \cdot (\|p_k - p_n\| - l_{k,n}) \tag{6}$$

where:

$$k = \text{the spring constant}$$
$$\|p_k - p_n\| = \text{the current Euclidean distance between points } p_k \text{ and } p_n$$

The spring constant influences only the scale, so its value can be assumed to be 1. The solution is obtained using a simulation approach (with discrete time) for a dynamic system described above—the positions of mass points are found through an iterative method. Damping is introduced in order to find a static solution (a system which minimizes energy) without oscillations. The energy E of a dynamic system is:

$$E = \sum_{i=1}^{n-1} \sum_{j=i+2}^{n} \frac{1}{2} \cdot k \cdot (\|p_k - p_n\| - l_{k,n})^2 \tag{7}$$

As a result, the projection which minimizes the total difference between the distance between points in the Cartesian coordinate system and the distance in a metric space is obtained. The points constitute projections of the distance graph vertices in 2D (or 3D) Cartesian coordinate system.

4.3 Experimental Results and Analysis

The example shows the structure of one of the biggest Polish news portals Onet (http://www.onet.pl). The distance graph contains over 5000 unique points (tokenized web pages) obtained while crawling these web sites starting from the home page. Figure 2 presents the projection of vertices of distance graph for the web site into 2D Cartesian coordinate system.

As can be seen in this figure, the distance graph contains a number of objects, i.e. individual web pages, that are relatively far from the agglomeration (concentration) area. Two examples of web pages from the agglomeration area are presented in Fig. 3, while the sample outliers are presented in Fig. 4. As can be noticed, these outlier web

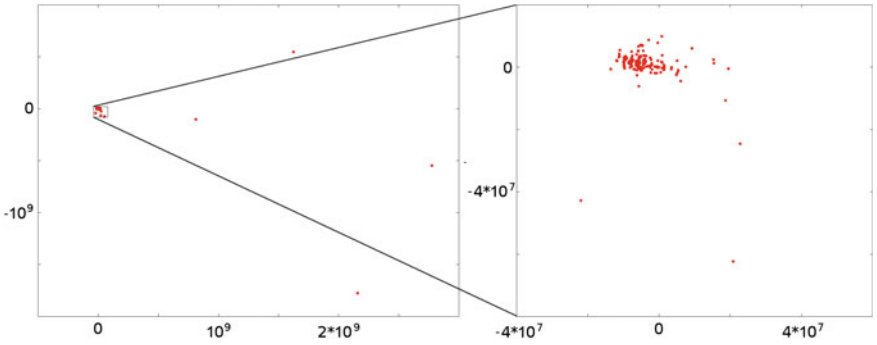


Fig. 2 2D projection of vertices of the distance graph for the web site. The *right* part of the image represents an enlarged part of the whole space (*left*)

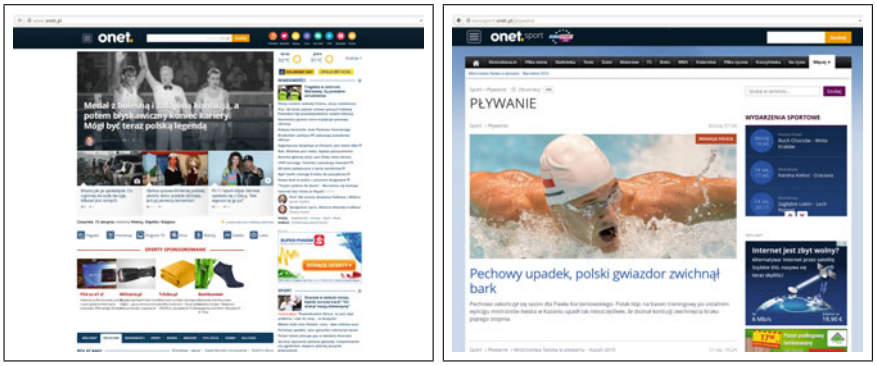


Fig. 3 Examples of typical web pages of the news portal from the primary agglomeration area

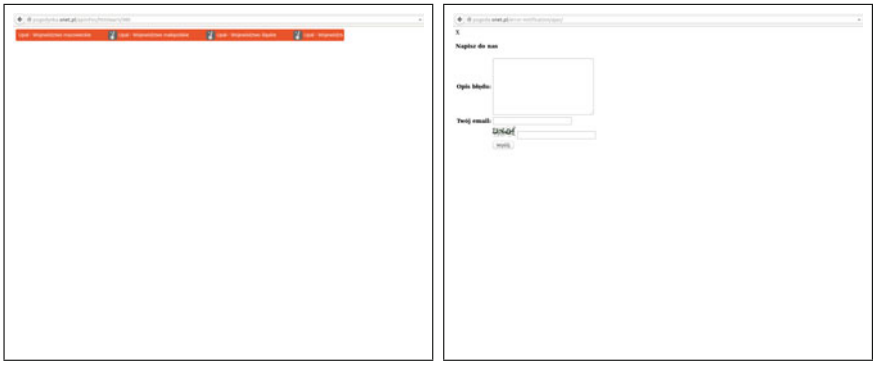


Fig. 4 Examples of some web pages far from the agglomeration area

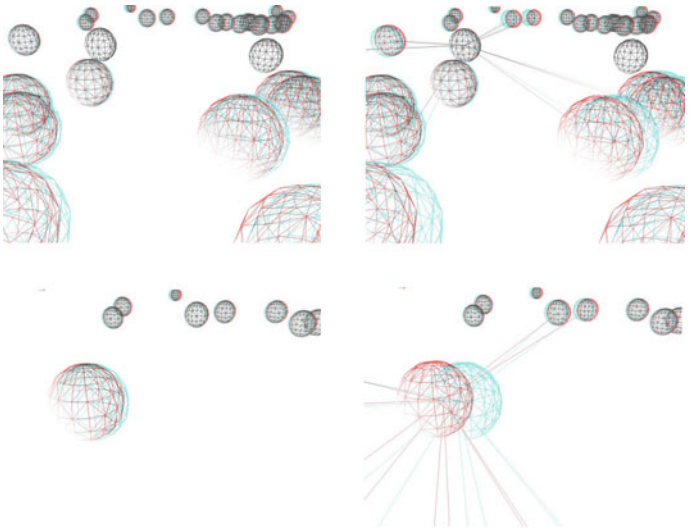


Fig. 5 An example of the visualization of a distance graph. Two different examples (*top* and *bottom*), both without the stereoscopic focus (*left*) and with the stereoscopic focus (*right*)

pages are quite different from a typical page. The fact that there are direct links to these outlier web pages likely denotes a mistake of the system administrator.

Although the projection of a distance graph into 2D Euclidean space (with the Cartesian coordinate system) in certain cases may be sufficient, the 3D projection offers many more possibilities of the structure interpretation on the basis of a distance graph.

To present a distance graph for a web site (described above) created in an iterative way, the stereoscopic focus can be set to the newly added vertex in a distance graph (i.e. the vertex representing the currently analysed web page during the process of web crawling). Edge cost value is defined as the edit distance between the two (tokenized) web pages. Figure 5 shows a fragment of a graph which includes a newly added node (which represents a currently analysed web page)—a 3D visualization without the stereoscopic focus (on the left) and with the stereoscopic focus applied (on the right).

5 Conclusions

This article focuses on a new method of presenting a graph-based data using stereoscopic vision. The method allows for drawing users' attention to specific elements or paths within the graph by utilizing an intentional discrepancy in the calculation of the nodes placement in the 3D space. Together with the proposed method of importance propagation over the graph's nodes, it smoothly draws observer's attention to

the particular fragment of the presented graph. The method was tested on a number of individuals, using a prototype platform with a specially designed 3D rendering engine developed by the authors. Multi-dimension graph's content was generated in many variants to investigate the effects of the attention drawing in various cases. The method could be extremely helpful when presented graph's nodes are in constant motion. This was found extremely useful in more complex graph visualizations during tests. Because the method does not use classical attention drawing measures (e.g. node attributes like colour or brightness), they can still be used for other purposes.

The effectiveness of attention drawing method naturally has certain limits. If the motion is too intensive or the distance changes too rapidly, the observer can lose "stereoscopic perception". It could happen in any kind of dynamic stereoscopic picture observations, but most likely when additional stereoscopic shifts are applied. Therefore, the strength of stereoscopic attention drawing assets must be slightly restrained if presentation dynamics is getting too violent.

References

1. Alper, B., Höllerer, T., Kuchera-Morin, J., Forbes, A.: Stereoscopic highlighting: 2D graph visualization on stereo displays. IEEE Trans. Vis. Comput. Graph. **17**(12), 2325–2333 (2011)
2. AlTarawneh, R., Bauer, J., Humayoun, S.R., Keller, P., Ebert, A.: The extended stereoscopic highlighting technique for node-link diagrams: An empirical study. In: Proceedings of the 14th IASTED International Conference on Computer Graphics and Imaging (CGIM 2013), Innsbruck, Austria (2013)
3. Andoni, A., Onak, K.: Approximating edit distance in near-linear time. SIAM J. Comput. **41**(6), 1635–1648 (2012)
4. Borg, I., Groenen, P.: Modern Multidimensional Scaling: Theory and Applications. Springer (2005)
5. Buss, S.R.: 3D Computer Graphics: A Mathematical Introduction with OpenGL. Cambridge University Press (2003)
6. Fauster, L., Wien, T.: Stereoscopic Techniques in Computer Graphics. Tu Wien (2007)
7. Greffard, N., Picarougne, F., Kuntz, P.: Beyond the classical monoscopic 3D in graph analytics: an experimental study of the impact of stereoscopy. In: 2014 IEEE VIS International Workshop on 3DVis (3DVis), pp. 19–24. IEEE (2014)
8. Kamada, T., Kawai, S.: An algorithm for drawing general undirected graphs. Inf. Process. Lett. **31**(1), 7–15 (1989)
9. Krebs, S.W., Zwiener, J.: Performance Optimisations in Stereoscopic Rendering for Depth Perception
10. Levenshtein, V.: Binary codes capable of correcting deletions and insertions and reversals. Soviet Phys. Doklady **10**(8), 707–710 (1966)
11. Lucca, G.D., Penta, M.D., Fasolino, A.: An approach to identify duplicated web pages. In: Proceedings of the International Computer Software and Applications Confernce (COMPSAC). pp. 481–486 (2002)
12. Lyes, T.: Review of stereo vision. Technical Report CSTN-155, Computer Science, Massey University, Albany, North Shore 102-904, Auckland, New Zealand (2011)
13. Schroeder, W., Martin, K., Lorensen, B.: An Object-Oriented Approach To 3D Graphics, vol. 429. Prentice Hall (1997)
14. Shapiro, L., Stockman, G.: Computer Vision. Prentice Hall (2001)
15. Torgerson, W.S.: Multidimensional scaling: I. Theory and method. Psychometrika **17**(4), 401–419 (1952)

16. Wagner, R.A., Fischer, M.J.: The string-to-string correction problem. J. Assoc. Comput. Mach. **21**, 168–173 (1974)
17. Watt, A.H., Watt, A.: 3D Computer Graphics, vol. 2. Addison-Wesley, Reading (2000)
18. Zachara, M., Pałka, D.: Information systems architecture and technology. In: Proceedings of 36th International Conference on Information Systems Architecture and Technology—ISAT 2015 – Part II, Chap. Comparison of Text-Similarity Metrics for the Purpose of Identifying Identical Web Pages During Automated Web Application Testing, pp. 25–35. Springer International Publishing (2016)

Part IV
Natural Language in Information Systems

Semi-automatic and Human-Aided Translation Evaluation Metric (HMEANT) for Polish Language in Re-speaking and MT Assessment

Krzysztof Wołk, Danijel Korzinek and Krzysztof Marasek

Abstract In this article we report the initial results of experiments using HMEANT metric (semi-automatic evaluation metric used for scoring translation quality by matching semantic role fillers) on the Polish language. The metric is evaluated in the task of Machine Translation (MT) and in re-speaking quality assessment. GUI-based annotation interface was developed and with this tool (https://github.com/krzwolk/HMEANT-metric-for-Polish) evaluation was conducted practically by not IT-related personnel. Reliability, correlation with automatic metrics, language independence and time costs were analysed as well. Role labelling and alignment using GUI interface were done by two annotators with no related background (they were only instructed for about 10 min). The results of our experiments showed high inter-annotator agreement as far as role labelling was concerned and a good correlation of the HMEANT metric with human judgements based on re-speaking evaluation.

Keywords HMEANT metric · HMEANT polish · Machine translation evaluation · Text quality assessment · Re-speaking evaluation

1 Introduction

Currently correct assessment of translation quality or text similarity is one of the most important tasks in related research areas. Machine translation (MT) evaluation history is quite old but it must be noted that most of the currently applied evaluation

K. Wołk (✉) · D. Korzinek · K. Marasek
Polish-Japanese Academy of Information Technology, ul. Koszykowa 86, 02-008 Warsaw, Poland
e-mail: kwolk@pja.edu.pl; kwolk@pjwstk.edu.pl

D. Korzinek
e-mail: danijel@pja.edu.pl

K. Marasek
e-mail: kmarasek@pja.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*, Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_21

approaches and metrics were developed recently. Automatic metrics like BLEU [1], NIST [2], TER [3] and METEOR [4] require only reference translation to be compared with MT output. This resulted in visible speed-up in MT research field. These metrics to some extent correlate with manual human evaluation and in most cases are language independent. The main problem in popular MT evaluation metrics is the fact that they analyse only lexical similarity and ignore the semantics of translation [5]. An interesting alternative approach was proposed by Snover et al. [6] and is known as HTER. This metric tries to measure the quality of translation in terms of post-editing actions. It was proved that this method correlates well with human equivalency judgments. However, the HTER metric is not commonly used in MT field because it requires high workload and time.

That is why in 2011 Lo and Wu [5] a family of metrics called MEANT was proposed. It proposes different evaluation approach. The idea is to measure how much an event structure of reference is preserved in translation. It utilizes shallow semantic parsing (MEANT metric) or human annotation (HMEANT) as a high standard.

In this research we implemented HMEANT for a new language (Polish) and we conducted evaluation on the usefulness of the new metric. The practical usage of HMEANT for the Polish language was analysed in accordance to criteria provided by Birch et al. [7]. Reliability in new language was measured as inter-annotator agreement (IAA) for individual stages during the evaluation task. Discriminatory power was estimated as the correlation of HMEANT rankings of our re-speaking results (obtained in manual human evaluation—NER and Reduction, automatic-evaluation—BLEU metric and the HMEANT) that was measured on the level of test sets. The language independence was analysed by collecting the problems of the original method and guidelines to them as described in Bojar and Wu [8] and Birch et al. [7]. Efficiency was studied as the workload cost of annotation task (average time required to evaluate translations) within HMEANT metric. Moreover, we choose unexperienced annotators to prove that semantic role labelling (SRL) does not require professional personnel. Being aware that some of the problems existing in HMEANT metric were already outlined [7, 8] and some improvements were already proposed, decision was made to conduct experiments with native HMEANT form. No changes to the metric itself were done, excepting the annotation interface enhancements that were made.

2 State of the Art

The first article on MEANT [5] proposed the semi-automatic evaluation approach. It required annotated event structure of two text sequences (reference and translation). The idea was to consider translation correct if it preserved shallow semantic (predicate-argument) structure of reference. Such structure was described in detail in [9]. In simple words the evaluation is conducted by probing events in the sentence (“Who did what to whom, when, where, why and how?”, etc.).

Such structures must be annotated and aligned across two translations. The founders of MEANT reported results of experiments, that utilized human annotation, semantic role labelling and an automatic shallow semantic parser. Their results showed that HMEANT metric correlates with human judgments at the value of 0.43 (Kendall tau, sentence level). This was a very close to the HTER correlation. In contrast BLEU has only 0.20. Also inter-annotator agreement (IAA) was analysed in two stages during the annotation process (role identification and role classification). The IAA ranged from 0.72 to 0.93 (good agreement).

For Czech-English translations MEANT and HMEANT metrics were used by Bojar and Wu [8]. Their experiments were conducted on a human evaluated set containing 40 translations from WMT12 conference. Sets were submitted by 13 different MT systems and each system was evaluated by exactly one annotator (inter-annotator agreement was not examined). HMEANT correlation against human evaluation was equal to 0.28 (much lower than the results of Lo and Wu [5].

In German-English translation Birch et al. [7] analysed HMEANT in accordance to four criteria, that address the usefulness of the metric in terms of reliability, efficiency, discriminatory power and language independence. The authors conducted experiments on evaluating 3 MT systems using a set of 214 sentences (142 in German and 72 in English). The IAA was divided into annotation and alignment steps. The results showed that the IAA for HMEANT was good enough at the first stages of the annotation but compounding effect of disagreements reduced the effective final IAA to 0.44 for German and to 0.59 for English. The efficiency of HMEANT was stated as reasonably good but it was not compared to other manual metrics.

3 Evaluation Using HMEANT

The annotation step in HMEANT has two stages semantic role labelling (SRL) and alignment. In the SRL phase annotators are asked to mark all the frames (a predicate and its roles) in reference and translated texts. In order to annotate a frame, it is necessary to mark the its predicate (a verb, but not a modal verb) and its arguments (role fillers—linked to that predicate). The role fillers are chosen from the inventory of 11 roles [5] (Table 1). Each role corresponds to a specific question about the entire frame.

Table 1 The role inventory

Who?	**What?**	**Whom?**
Agent	Patient	Benefactive
When?	**Where?**	**Why?**
Temporal	Locative	Purpose
How?		
Manner, degree, negation, modal, other		

Secondly, the annotators need to align the elements of frames. They must link both actions and roles, and mark them as "Correct" or "Partially Correct" (depending on equivalency in their meaning). In this research we used the original guidelines for the SRL and alignment described in [5].

3.1 HMEANT Calculation

Having the annotation step completed the HMEANT score can be calculated as the F-score from the counts of matches of predicates and corresponding role fillers [5]. Predicates (together with roles) not having correct matches are not taken into account. The HMEANT model is defined as follows:

$\#F_i$—number of correct role fillers for predicate i in machine translation.
$\#F_i$ *(Partial)*—number of partially correct role fillers for predicate i in machine translation.
$\#MT_i$, $\#REF_i$—total number of role fillers in machine translation or reference for predicate i.
N_{mt}, N_{ref}—total number of predicates in MT or reference.
W—weight of the partial match (0.5 in the uniform model).

$$P = \sum_{matched\ i} \frac{\#F_i}{\#MT_i}$$

$$R = \sum_{matched\ i} \frac{\#F_i}{\#REF_i}$$

$$P_{part} = \sum_{matched\ i} \frac{\#F_i(partial)}{\#MT_i}$$

$$R_{part} = \sum_{matched\ i} \frac{\#F_i(partial)}{\#REF_i}$$

$$P_{total} = \frac{P + w * P_{part}}{N_{mt}}$$

$$R_{total} = \frac{R + w * R}{N_{ref}}$$

$$HMEANT = \frac{2 * P_{total} * R_{total}}{P_{total} + R_{total}}$$

3.2 Calculation of the Inter Annotator Agreement

In accordance to Lo and Wu [5] and Birch et al. [7] the IAA was studied as well. It is defined as F1-measure in which one of the annotators is considered to be a gold standard as follows:

$$IAA = \frac{2 * P * R}{P + R}$$

where P is precision (number of labels [roles, predicates or alignments], that were matched between the annotators). Recall (R) is defined as quantity of matched labels divided by total quantity of labels. In accordance to Birch et al. [7] only exact word span matches is considered. The stages of the annotation process described in [7] were adopted as well (role identification, role classification, action identification, role alignment, action alignment). Disagreements were analysed by calculating the IAA for each stage separately.

4 Human Calculated NER Metric

The NER model [10] is a simple extension of the word accuracy metric adapted specifically for measuring the accuracy of subtitles. It is one of two measures that is of particular importance for media providers (television companies, movie distributors, etc.), the other one being the reduction rate. Generally, the aim of good subtitles is to reduce the length of written text as much as possible (in order to preserve space on screen and make it easier to read) while maintaining an almost perfect accuracy (usually above 98 %).

Since we are dealing with paraphrasing, it is very difficult to perform accurate measurements by comparing the text only. The NER model gets around this problem by counting errors using a simple formula, which inspired its name:

$$\text{NER accuracy} = \frac{\text{N} - \text{E} - \text{R}}{\text{N}} \times 100\,\%$$

where N is the number of analysed tokens (usually also includes punctuation), E is the number of errors performed by the re-speaker, and R is the number of errors performed by the ASR system (on re-speaker's speech). Additionally, the errors in E are weighted: 0.25 for minor errors, 0.5 for normal and 1 for serious errors. There are user-friendly tools available for making these assessments and obviously there may be a certain level of human bias involved in these estimates. Nevertheless, all the decisions are thoroughly checked and explainable using this method, which makes it one of the most popular techniques for subtitle quality assessment used by many regulatory bodies worldwide.

5 Experimental Setup

The data used in the experiments described in this paper was collected during a study performed in the Institute of Applied Linguistics at the University of Warsaw [10]. This, still ongoing, study aims to determine the relevant features of good re-speakers and their technique. One of the obvious measures is naturally the quality of re-speaking discussed in this paper.

The subjects were studied in several hour sessions, where they had to perform various tasks and psychological exams. The final test was to do actual re-speaking of pre-recorded material. This was simultaneously recorded and recognized by a commercial, off-the-shelf ASR suite. The software was moderately adapted to the re-speaker during a several hour session, a few weeks before the test.

The materials included four different 5 min segments in the speaker's native language and one in English (where the task was also translation). The recordings were additionally transcribed by a human, to convey exactly what the person said. The final dataset contains three sets of transcriptions:

1. the transcription of the original recorded material.
2. the transcription of the re-speaker transcribed by a human (used in the evaluation).
3. the output of the ASR recognizing the re-speaker.

The Table 2 contains results of human evaluation using NER and Reduction (goal is to have as big NER and Reduction as possible at the same time) and semi-automatic using HMEANT metric.

We asked to participate two annotators with no linguistic background. Every annotator evaluated exactly 15 re-spoken samples each of them contained 40 sentences that were supposed to be annotated by them.

6 Results and Conclusions

Backward linear regression has been used in analysis for the following reasons:

1. The data is linear (correlation analysis present linear relation).
2. The data is ratio level data, thus good for linear regression.
3. The regression analysis provides more than one regression results, thus the best variables can be found.
4. Reliable variables are extracted by excluding irrelevant variables from the 1st to the last stage (give the most reliable variables).

For this analysis, the Standardized Coefficients would have the strength of relationship extracted from the Unstandardized Coefficients. The sig (p-value) was judged with the alpha (0.05) to investigate the most significant variables that

Table 2 Results of human evaluation

NER	HMEANT	REDUCTION
93.61	0.6016	22.77
94.09	0.6626	14.33
94.78	0.871	9.31
95.96	0.9395	2.71
94.77	0.6854	10.15
97.01	0.8958	3.89
95.83	0.7407	8.63
87.56	0.4621	25.21
85.76	0.4528	28.09
93.98	0.7148	8.63
95.79	0.5994	11.68
94.77	0.4641	19.12
97.41	0.9253	0.34
93.76	0.9529	5.75
93.89	0.8086	4.4
93.97	0.8194	7.45
92.74	0.725	8.46
88.07	0.6437	25.21
86.68	0.2976	32.66
84.36	0.4148	52.12
95.01	0.719	11.84
92.23	0.6244	11.84
93.11	0.5676	27.58
95.36	0.796	6.09
89.97	0.589	21.15
94.95	0.7975	5.41
95.75	0.8302	7.45
98.69	0.8869	4.23
89.37	0.5852	−6.6
93.73	0.8144	2.2

explain the NER metrics. The Adjusted R-square (R^2) would provide the information, how much the variables are explaining the NER variances.

Table 3 represents the regression summary for NER with HMEANT and REDUCTION metrics. Here, at first, the model has both HMEANT and REDUCTION metrics and HMEANT is the significant predictors of NER. Additionally, REDUCTION is the insignificant metric, thus it has been removed for the second model. In the second model only HMEANT ($p = 0.000$) is significant as it has p-value below the alpha (0.05). In this case, the final model (2nd model) shows that, the B-value for HMEANT is 0.162, indicating one-unit increase in HMEANT value would bring 0.162 times increase in NER value, higher beta value (B) indicates strong relationship. While the constant for this case 81.824 (higher constant

Table 3 Regression result summary for NER and HMEANT, REDUCTION metrics

Model		Unstandardized coefficients		Standardized coefficients	T	Sig.	Adjusted R-square
		B	Std. error	Beta			
1st model	(Constant)	81.814	1.825		44.838	0.000	0.603
	HMEANT	0.156	0.026	0.744	5.925	0.000	
	REDUCTION	0.030	0.038	0.102	0.810	**0.425**	
2nd model	(Constant)	**81.824**	1.813		45.123	0.000	0.594
	HMEANT	**0.162**	0.025	0.770	6.395	**0.000**	
a. Dependent Variable: NER							

value indicates more error in the model). The Regression overall explains 59.4 % variability of NER values, which is statistically acceptable.

From the two models it has been found that HMEANT is significant predictor of NER (which means it correlates with human judgments correctly). The regression equation that can compute the value of NER based on HMEANT statistical metrics as:

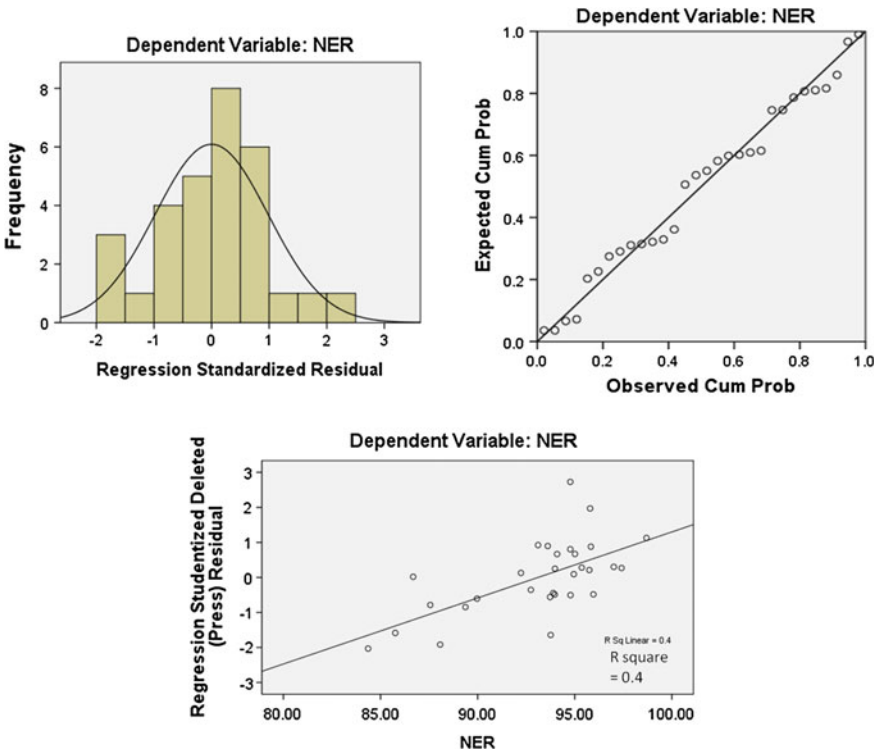


Fig. 1 Average sentence lengths

$$\mathrm{NER} = 81.82 + 0.162 * \mathrm{HMEANT}$$

Moving forward, regression residual plots from the above regression for the significant metrics are presented in Fig. 1. The histogram and normality plots show the distribution of residual and scatter plot shows relation between dependent metric with regression residual. The closer the dots in the plot to the regression line the better the R-square value and the better the relationship.

Figure 1, shows the regression residual analysis for HMEANT, and it is clear that the histogram and normality plot show the residuals are distributed normally.

Lastly we analysed the IAA dividing it into annotation and alignment steps. The agreement in annotation was quite high equal to 0.87 but in alignment step the agreement was 0.63. Most likely because of the subjective human feelings about the meaning. In addition, we measured that time average required for an inexperienced annotation to evaluate one 40 sentence long texts was about 27 min.

Acknowledgments This research was supported by Polish-Japanese Academy of Information Technology statutory resources (ST/MUL/2016) and resources for young researchers.

References

1. Papineni, K., et al.: BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
2. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 138–145. Morgan Kaufmann Publishers Inc. (2002)
3. Koehn, P., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. Association for Computational Linguistics (2007)
4. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
5. Lo, C., Wu, D.: MEANT: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 220–229. Association for Computational Linguistics (2011)
6. Snover, M., et al.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231 (2006)
7. Birch, A., et al.: The feasibility of HMEANT as a human MT evaluation metric. In: Proceedings of the Eighth Workshop on Statistical Machine Translation, pp. 52–61 (2013)
8. Bojar, O., Wu, D.: Towards a predicate-argument evaluation for MT. In: Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 30–38. Association for Computational Linguistics (2012)
9. Pradhan, S., et al.: Shallow Semantic Parsing using Support Vector Machines. HLT-NAACL, pp. 233–240 (2004)
10. Wołk, K., Korzinek, D.: Comparison and Adaptation of Automatic Evaluation Metrics for Quality Assessment of Re-Speaking. arXiv:1601.02789 (2016)

Analysis of Complexity Between Spoken and Written Language for Statistical Machine Translation in West-Slavic Group

Agnieszka Wołk, Krzysztof Wołk and Krzysztof Marasek

Abstract The multilingual nature of the world makes translation a crucial requirement today. Within this research we apply state of the art statistical machine translation techniques to the West-Slavic languages group. We do West-Slavic languages classification and choose Polish as a representative candidate for our research. The experiments are conducted on written and spoken texts, which characteristics are defined as well. The machine translation systems are trained within West-Slavic group as well as into English. Translation systems and data sets are analyzed, prepared and adapted for the needs of West-Slavic—* translation. To evaluate the effects of different preparations on translation results, we conducted experiments and used the BLEU, NIST and TER metrics. By defining proper translation parameters to morphologically rich languages we improve the translation quality and draw the conclusions.

Keywords Statistical machine translation · Complexity · West-Slavic

1 Introduction

When implementing Statistical Machine Translation systems (SMT) it is necessary to deal with many problems to achieve high quality translations. These problems include the need to align parallel texts in language pairs and clean parallel corpora to remove errors. This is especially true for real-world corpora developed from text harvested from the vast data available in the Internet. Out-Of-Vocabulary

A. Wołk · K. Wołk (✉) · K. Marasek
Polish-Japanese Academy of Information Technology, ul. Koszykowa 86,
02-008 Warsaw, Poland
e-mail: kwolk@pja.edu.pl; kwolk@pjwstk.edu.pl

A. Wołk
e-mail: awolk@pja.edu.pl

K. Marasek
e-mail: kmarasek@pja.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_22

(OOV) words must also be handled, as they are inevitable in real-world texts [1]. The lack of enough parallel corpora is another significant challenge for SMT. Since the approach is statistical in nature, a significant amount of quality language pair data is needed to improve translation accuracy. In addition, very general translation systems that work in a general text domain have accuracy problems in specific domains. SMT systems are more accurate on corpora from a domain that is not too wide.

Another problem not yet addressed is adaptation of machine translation techniques for the needs of the specific language or event kind of a text. Such adaptation would be very beneficial especially for very diverse languages like e.g., Slavic-English. In this research we focus on very complex West-Slavic languages, which represent a serious challenge to any SMT system. They differ from other languages similar to English in grammar full complicated rules and elements, together with a big vocabulary (due to complex declension). The main reasons for its complexity are more cases, genders, animate and inanimate nouns, adjectives agreements with nouns in terms of gender, case, number and a lot of words borrowed from other languages which are often inflected similarly to those of West-Slavic origin. This greatly affects the data and data structure required for statistical models of translation. The lack of available and appropriate resources required for data input to SMT systems presents another problem especially when vocabularies are disproportionate (which is the case).

In our research we conduct experiments that try to overcome those disproportion and improve the baseline systems based on Moses SMT state of the art configuration. We conduct research on Polish language as a representative candidate from West-Slavic group and try to translate it between English an Czech. In addition, we do comparison of translation between spoken and written texts because in recent research [2], the authors analysed the differences between spoken language based on lectures and written texts based on written news for machine translation. They showed that there are meaningful differences between those two text genres and that particular statistical modeling strategies should be independently applied to each translation task.

The article is structured as follows. Section 2 describes West-Slavic languages and main differences between Polish and English languages, Sect. 3 contains comparison of spoken and written texts used in this research. Section 4 describes our machine translation systems and evaluation methods. Lastly we draw conclusions in the Sect. 5.

2 West-Slavic Languages Group

West-Slavic languages are part of Slavic languages (alongside East- and South-Slavic), which are used by approximately 56 million people in the Central Europe. West-Slavic languages split in: Lechitic languages (Polish, Silesian, and Kashubian,

the latter two being dialects of Polish), Sorbian languages (Upper and Lower), Czech-Slovak (Czech and Slovak) [3].

Approximately, a native speaker of Polish uses in his everyday language not much over a 12,000 words actively, whereas passively about 30,000 words. However, research of philologist from Bialystok University shows that knowledge of 1,200 most commonly used Polish words is enough to communicate effectively [4].

Kashubian, is considered either a separate language or a dialect of Polish. Often, indirect position is adopted, and terms such as "language" or "dialect" are avoided, and instead Kashubian is classified as an "ethnolect". In law, Kashubian is regarded as a regional language. Every day, around 108,000 Polish people use Kashubian [5].

Silesian ethnolect is a set of Silesian jargons, possibly mixing into multiple dialects. It is used by indigenous inhabitants of Upper Silesia and a small population of Lower Silesia. Languages like literary Polish, Czech (particularly from Moravian, formerly functioning as a separate language), German (mostly Germanic dialect of Silesia), and partly Slovak played a huge role in the process of formation of Silesian ethnolect, as we can trace many loanwords from those languages [6].

The Sorbian languages is a group consisting of two closely related West Slavic languages: Sorbian and Lower Sorbian. They are used by the Sorbs (totalling approximately 150,000 people, most of whom speak only German) in Lausitz in eastern Germany. Both languages are endangered Sorbian languages (especially Lower Sorbian; the number of active users is estimated at several thousand; mostly used only by the older generation, while the Upper Lusatian's users total approximately 55,000 people). Lower Sorbian shares more similarities to Polish, while Sorbian with Czech and Slovak.

Czech derives from Proto-Indo-European through the Proto-Slavic. It developed orally in tenth century and its first artefacts date back to the thirteenth century. Late Middle Ages was a period of flourishing Czech and its strong influence on other languages (including Polish).

Silesian and Kashubian are very similar to Polish in terms of grammar. Both Upper and Lower Sorbian have the dual for nouns, pronouns, adjectives and verbs; very few known living Indo-European languages retain this feature as a productive aspect of the grammar. For example, the word ruka is used for one hand, ruce for two hands, and ruki for more than two hands [6].

There are four grammatical genders in Slovak language: animate masculine, inanimate masculine, feminine and neuter. In popular description, the first two genders are often covered under common masculine gender. There is the singular and the plural numbers. The following morphological cases are present in Slovak: the nominative case answers the question Who/What; the genitive case (of whom); the dative case (to whom); the accusative case (whom); the locative case (used after the prepositions); the instrumental case (by means of whom).

An adjective's, pronoun's, and to some extent also the number's person, gender, and case decline according to the noun. An adjective always precedes the corresponding noun. The comparative is formed by replacing the adjective ending -ý/y/i/í by -ejší or -ší. There are exact rules for the choice between these two endings and

there are several irregular comparatives. The superlative is formed as follows: naj +comparative. The comparative and superlative of adverbs (which end in -o, -e or -y in the basic form) is formed by simply replacing the -(ej)ší from the adjective by -(ej)šie. The verb (predicate) agrees in person and number with its subject [4].

Czech is very similar to Polish and Slovak. The most recognizable thing in Czech, which occurs also in Polish is fact that one word is often sufficient to express what English can only achieve by using multiple words.

2.1 Differences Between Polish and English Languages

In general, Polish and English differ in syntax and grammar. English is a positional language, which means that the syntactic order (the order of words in a sentence) plays a very important role, particularly due to the limited inflection of words (e.g., lack of declension endings). Sometimes, the position of a word in a sentence is the only indicator of the sentence's meaning. In a Polish sentence, a thought can be expressed using several different word orderings, which is not possible in English. For example, the sentence "I bought myself a new car." can be written in Polish as "Kupiłem sobie nowy samochód.", or "Nowy samochód sobie kupiłem.", or "Sobie kupiłem nowy samochód.", or "Samochód nowy sobie kupiłem." The only exception is when the subject and the object are in the same clause and the context is the only indication which is the object and which is subject. For example, "Mysz liże kość. (A mouse is licking a bone.)" and "Kość liże mysz. (A bone is licking a mouse).".

Differences in potential sentence word order make the translation process more complex, especially when using a phrase-model with no additional lexical information [7]. In addition, in Polish it is not necessary to use the operator, because the Polish form of a verb always contains information about the subject of a sentence. For example, the sentence "On jutro jedzie na wakacje." is equivalent to the Polish "Jutro jedzie na wakacje." and would be translated as "He is going on vacation tomorrow." [8].

In the Polish language, the plural formation is not made by adding the letter "s" as a suffix to a word, but rather each word has its own plural variant (e.g., "pies—psy", "artysta—artyści", etc.). Additionally, prefixes before nouns like "a", "an", "the", do not exist in Polish (e.g., "a cat—kot", "an apple—jabłko", etc.) [8].

The Polish language has only three tenses (present, past, and future). However, it must be noted that the only indication whether an action has ended is an aspect. For example, "Robiłem pranie." Would be translated as "I have been doing laundry", but "Zrobiłem pranie." as "I have done laundry", or "płakać—wypłakać" as "cry—cry out" [8].

The gender of a noun in English does not have any effect on the form of a verb, but it does in Polish. For example, "Zrobił to. – He has done it.", "Zrobiła to. – She has done it.", "lekarz/lekarka—doctor", "uczeń/uczennica—student", etc. [8].

As a result of this complexity, progress in the development of SMT systems for West-Slavic languages has been substantially slower than for other languages. On the other hand, excellent translation systems have been developed for many popular languages.

3 Spoken Versus Written Language

The differences between speech and text within the context of the literature should also be clarified. Chong [9] pointed out that writing and speech differ considerably in both function and style. Writing tends towards greater precision and detail, whilst speech is often punctuated with repetition and includes prosody, which writing does not possess, to further convey intent and tone beyond the meaning of the words themselves.

According to William Bright [10], spoken language consists of two basic units: Phonemes, units of sound, (that are themselves meaningless) are combined into morphemes, which are meaningful (e.g., the phonemes /b/, /i/, and /t/ form the word "bit"). Contrary alphabetic scripts work in similar way. In a different type of script, the basic unit corresponds to a spoken syllable. In logographic script (e.g., Chinese), each character corresponds to an entire morpheme, which is usually a word [10].

It is possible to convey the same messages in either speech or writing, but spoken language typically conveys more explicit information than writing. The spoken and written forms of a given language tend to correspond to one or more levels and may influence each other (e.g., "through" is spoken as "thru").

In addition, writing can be perceived as colder, or more impersonal, than speech. Spoken languages have dialects varying across geographical areas and social groups. Communication may be formal or casual. In literate societies, writing may be associated with a formal style and speech with a more casual style. Using speech requires simplification, as the average adult can read around 300 words per minute, but the same person would be able to follow only 150–200 spoken words in the same amount of time [11]. That is why speech is usually clearer and more constrained.

The punctuation and layout of written text do not have any spoken equivalent. But it must be noted that some forms of written language (e.g., instant messages or emails) are closer to spoken language. On the other hand, spoken language tends to be rich in repetition, incomplete sentences, corrections, and interruptions [12].

When using written texts, it is not possible to receive immediate feedback from the readers. Therefore, it is not possible to rely on context to clarify things. There is more need to explain things clearly and unambiguously than in speech, which is usually a dynamic interaction between two or more people. Context, situation, and shared knowledge play a major role in their communication. It allows us to leave information either unsaid or indirectly implied [12].

4 Machine Translation

Moses is a tool environment for statistical machine translation that enables users to train translation models for any two languages. This implementation of the statistical approach to machine translation is currently the dominant approach in this field of research [13].

4.1 Baseline System

The baseline system testing was done using the Moses open source SMT toolkit with its Experiment Management System (EMS) [13]. The SRI Language Modeling Toolkit (SRILM) [14] with an interpolated version of the Kneser-Ney discounting (interpolate –unk –kndiscount) was used for 5-gram language model training. The MGIZA++ tool was used for word and phrase alignment. KenLM [15] was used to binarize (transform features of a text entity into vectors of numbers) the language model, with the lexical reordering set to the msd-bidirectional-fe model [16]. The symmetrization method was set to grow-diag-final-and for word alignment processing [13].

4.2 Parameter Adaptation

We raised our score in PL-* experiments through changing the language model order from 5 to 6 and changed the discounting method from Kneser-Ney to Witten-Bell. In the training part, we changed the lexicalized reordering method from msd-bidirectional-fe to hier-mslr-bidirectional-fe. The system was also enriched with Operation Sequence Model (OSM) [17]. The motivation for OSM is that it provides phrase-based SMT models the ability to memorize dependencies and lexical triggers, it can search for any possible reordering, and it has a robust search mechanism. What is more we used Compound Splitting feature [13]. Tuning was done using MERT tool with batch-mira feature and n-best list size was changed from 100 to 150.

Because of a much bigger dictionary, the translation from EN to PL is significantly more complicated. We determined that not using lower casing, changing maximum sentence length to 85, maximum phrase length to 7 improves the BLEU score. Additionally, we set the language model order from 5 to 6 and changed the discounting method from Kneser-Ney to Witten-Bell. In the training part, we changed the lexicalized reordering method from msd-bidirectional-fe to tgttosrc. The system was also enriched with Operation Sequence Model (OSM). What is

more we used Compound Splitting feature and we did punctuation normalization. Tuning was done using MERT tool with batch-mira feature and n-best list size was changed from 100 to 150. Training a hierarchical phrase-based translation model also improved results in this translation scenario [18].

In PL-CS experiments it was necessary to adjust parameters so that they suit translation between two morphologically rich languages. We changed language model order to 8 which in contrast with PL-EN translation produced positive results. The maximum sentence length was changed to 100 from 80 (in PL-EN exceeding 85 did not rise the scores). The lexicalized reordering method was changed to wbe-msd-bidirectional-fe, in order to use word-based data extraction. In tuning the n-best list size was raised to 200 because there were many possible candidates to be choose in that language pair.

5 Evaluation

Metrics are necessary to measure the quality of translations produced by the SMT systems. For this, various automated metrics are available to compare SMT translations to high quality human translations. Since each human translator produces a translation with different word choices and orders, the best metrics measure SMT output against multiple reference human translations. For scoring purposes we used four well-known metrics that show high correlation with human judgments. Among the commonly used SMT metrics are: Bilingual Evaluation Understudy (BLEU), the U.S. National Institute of Standards & Technology (NIST) metric, the Metric for Evaluation of Translation with Explicit Ordering (METEOR), and Translation Error Rate (TER). According to Koehn, BLEU [16] uses textual phrases of varying length to match SMT and reference translations. Scoring of this metric is determined by the weighted averages of those matches [19]. To encourage infrequently used word translation, the NIST [19] metric scores the translation of such words higher and uses the arithmetic mean of the n-gram matches. Smaller differences in phrase length incur a smaller brevity penalty. This metric has shown advantages over the BLEU metric. The METEOR [19] metric also changes the brevity penalty used by BLEU, uses the arithmetic mean like NIST, and considers matches in word order through examination of higher order n-grams. These changes increase score based on recall. It also considers best matches against multiple reference translations when evaluating the SMT output. TER [20] compares the SMT and reference translations to determine the minimum number of edits a human would need to make for the translations to be equivalent in both fluency and semantics. The closest match to a reference translation is used in this metric. There are several types of edits considered: word deletion, word insertion, word order, word substitution, and phrase order.

Table 1 Corpora statistics

Corpus	Language pair	Number of sentences	Number of unique tokens		Avg. sentence length
			PL	FOREIGN	
EUB	PL-EN	537,941	346,417	207,563	23
	PL-CS	435,376	300,897	299,402	22
QED	PL-EN	331,028	168,961	60,839	12
	PL-CS	475,621	172,891	159,325	14

Table 2 Translation of written language

SYSTEM	Direction	BLEU	NIST	TER	METEOR
BASE	PL → CS	28.93	5.90	67.75	44.39
BEST	PL → CS	29.76	5.98	65.91	45.84
BASE	PL → EN	31.63	6.27	66.93	53.24
BEST	PL → EN	32.81	6.47	65.12	53.92
BASE	CS → PL	26.32	5.11	73.12	42.91
BEST	CS → PL	27.41	5.66	70.63	43.19
BASE	EN → PL	22.18	5.07	74.93	39.13
BEST	EN → PL	23.43	5.33	72.18	41.24

6 Results and Conclusions

The experiment results were gathered in Tables 2, 3 and 4. BASE stands for baseline systems settings and BEST for translation systems with modified training settings (in accordance to Sect. 4.2). Table 2 contains translation results for written texts and Table 3 for spoken language. Decision was made to use EU Bookshop[1] (EUB) document-based corpus as an example of written language and the QCRI Educational Domain Corpus[2] (QED) (open multilingual collection of subtitles for educational videos and lectures). Both corpora are comparable examples of spoken and written language. The corpora specification is showed in Table 1.

The results presented in Table 1 confirm statements from the Sect. 3. There is a big dictionary gap between West-Slavic languages and English that is not present within the West-Slavic group itself. We can also observe big difference in sentence lengths between written and spoken language.

As presented in Tables 2 and 3 it can be observed that usually translation between West-Slavic and English language gives better results that translation within the West-Slavic group. Most likely reason is the morphological richness in this group which produces many to many word and phrase mappings making it difficult statistically choose correct ones. In addition, translation of spoken language

[1] http://bookshop.europa.eu.

[2] http://alt.qcri.org/resources/qedcorpus/.

Table 3 Translation of spoken language

SYSTEM	Direction	BLEU	NIST	TER	METEOR
BASE	PL → CS	7.11	3.27	79.96	29.23
BEST	PL → CS	8.76	3.95	74.12	32.34
BASE	PL → EN	15.28	5.08	68.45	48.49
BEST	PL → EN	15.64	5.11	68.24	48.75
BASE	CS → PL	6.77	3.12	76.35	29.89
BEST	CS → PL	7.43	3.51	75.76	31.73
BASE	EN → PL	7.76	3.55	79.32	31.04
BEST	EN → PL	8.22	3.78	77.24	32.33

Table 4 Translation with EN as pivot language

CORPUS	Direction	BLEU	NIST	TER	METEOR
EUB	PL → CS	28.93	5.90	67.75	44.39
EUB	PL → EN → CS	26.12	4.77	72.03	41.95
QED	PL → CS	7.11	3.27	79.96	29.23
QED	PL → EN → CS	9.23	4.18	76.33	41.42

produces much worse results that those of written texts. Some of this difference is obviously the gap between the size of training data but high level of formalisms in written language structures seems to produce the biggest impact. What is more, those formal language structures make it possible to overcome disparities between languages (the difference in scores between PL-CS and PL-EN translation is small, such phenomena is not replicated in spoken language). Because it was observed that in case of speech there is big difference in translation quality between PL-EN and PL-CS it was decided to conduct addition experiment on translation from PL to CS using EN as a pivot language. Such experiment is showed in Table 4. Such idea produced positive results in spoken language translation but negative in written. Most likely reason for this is the fact that in written language translation quality was already similar for both pairs.

Acknowledgments This research was supported by Polish-Japanese Academy of Information Technology statutory resources (ST/MUL/2016), resources for young researchers at PJATK and CLARIN ERIC research program.

References

1. Mohammadi, M.; Ghasemaghaee, N.: Building bilingual parallel corpora based on wikipedia. In: 2010 Second International Conference on Computer Engineering and Applications (ICCEA), pp. 264–268. IEEE (2010)
2. Ruiz, N., Federico, M.: Complexity of spoken versus written language for machine translation. In: Proceedings of the 17th Annual Conference of the European Association for Machine Translation, pp. 173–180 (2014)

3. Dalewska-Greń, H.: Języki słowiańskie. Wydawn, Naukowe PWN (1997)
4. Stieber, Z.: Zarys gramatyki porównawczej języków słowiańskich Wydawn. Naukowe PWN (2005)
5. Oczkowa, B., Szczepańska, E., Kwoka T.: Słowiańskie języki literackie. Wydawnictwo Uniwersytetu Jagiellońskiego (2011)
6. Języki zachodniosłowiańskie last modified October 16 2015. https://pl.wikipedia.org/wiki/J%C4%99zyki_zachodnios%C5%82owia%C5%84skie
7. Wołk, K., Marasek, K.: Polish–English speech statistical machine translation systems for the IWSLT 2013. In: Proceedings of the 10th International Workshop on Spoken Language Translation, Heidelberg, Germany, pp. 113–119 (2013)
8. Swan, O.E.: Polish Grammar in a Nutshell (2003)
9. Choong C.: The Difference between Written and Spoken English. Assignment Unit 1 A in fulfillment of Graduate Diploma in English (2014)
10. Daniels, P. T., Bright, W.: The World's Writing Systems. Oxford University Press (1996)
11. Coleman, J., A speech is not an essay. Harv. Bus. Rev. (2014)
12. Ager, S.: Differences between writing and speech, Omniglot—the online encyclopedia of writing systems and languages. http://www.omniglot.com/writing/writingvspeech.htm. Accessed 8 Aug 2013
13. Koehn, P., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. Association for Computational Linguistics (2007)
14. Stolcke, A., et al.: SRILM-an extensible language modeling toolkit. In: INTERSPEECH (2002)
15. Heafield, K.: KenLM: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 187–197. Association for Computational Linguistics (2011)
16. Gao, Q., Vogel, S.: Parallel implementations of word alignment tool. In: Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pp. 49–57. Association for Computational Linguistics (2008)
17. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features, pp. 137–142. Springer, Berlin, Heidelberg (1998)
18. Tiedemann, J.: News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In: Recent Advances in Natural Language Processing, pp. 237–248 (2009)
19. Wu, D., Fung, P.: Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In: Natural Language Processing–IJCNLP 2005, pp. 257–268. Springer, Berlin, Heidelberg (2005)
20. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. LREC **2012**, 2214–2218 (2012)

Interactive Gradually Generating Relevance Query Refinement Under the Human-Mediated Scenario in Multilingual Settings

Jolanta Mizera-Pietraszko and Aleksander Zgrzywa

Abstract As opposed to query modelling, relevance generating interactive query refinement (QR) is a technique aimed at exploiting syntax variations of gradually extended, being removed or replaced with some other keywords query, which depending on the factors like e.g. the information resource, the database structure, or the keyword alignment, facilitates significantly the searching process. Therefore our motivation is to explore the dynamism of the precision trend depended upon the factors analyzed. For a couple of language pairs which constitute multilingual settings, we develop a user-centred framework that imposes distributed search optimization. Our data set contains variety of query types submitted to some translingual distributed search systems that perform a number of syntax-based indexing. We construct a dynamism of precision elevation trend that indicates what factors intensify the relevance set of the system responses from a perspective of the user's information need.

Keywords Query refinement · Multilingual information retrieval · Distributed search

1 Introduction

Successive query formulation is one of the most formidable challenges for the users, specifically for the multilingual purposes, as it requires the knowledge about the prediction of the keywords that are matched by the documents. While

J. Mizera-Pietraszko (✉)
Institute of Mathematics and Informatics, Opole University, Opole, Poland
e-mail: jmizera@math.uni.opole.pl

A. Zgrzywa
Department of Information Systems, Wroclaw University of Science and Technology, Wrocław, Poland
e-mail: aleksander.zgrzywa@pwr.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_23

attempting to access very large databases any ambiguous query keyword produces a long list of irrelevant results taking a long time to sift through.

Multilingual search represents a high technology evidence when compared to bilingual retrieval, let alone monolingual search tasks. Though, due to the system architecture, a user does not have an opportunity for choice of the target language. Some reasons, the most users are unaware of the search modes other than free text, while narrowing the queries by excluding some keywords closely related to the topic, or alternatively searching only some URL links, enables more efficient exploration of the databases accessible for the particular search system.

In the approach presented, the starting point is that by submitting queries, the user systematically broadens the knowledge in the field of his or her interest, which results in entering more precise subsequent queries.

Alternatively while refining the query some of the keywords are removed and gradually replaced by those producing in Human-Mediated Scenario (HMS) more and more relevant results from the perspective of the user's need.

The remainder of this paper is organized as follows: at first we introduce the general concept of our work that is the framework of the experiment, then, we present the state-of-the-art technology, describe the multilingual setting and the interactive gradually generating relevance query refinement to move on to the analysis of dynamism of the precision elevation trend. In section three we show how to generate relevance under HMS. The next section introduces the metric of generative relevance. Finally, we construct the precision trend based on our results. In the last section we discuss the results and relate to the conclusion for the further research.

2 Literature Overview

Since relevance represents the ranking of search responses adequacy to the query entered by the user whose need is usually fuzzy, especially at the first attempt of interaction with the system, quantifying its score by making the use of the measures other than precision and recall with the aim to be found even more precise, still seems to be quite challenging despite the decades of intensive research in the area.

Relevance-generating query refinement relies on the user's gradually broadening knowledge acquired from browsing the Web, in particular the Deep Web oriented towards the information routinely missed by the standard search systems produced by the authors who hope to benefit from their anonymity in the Web [1–4].

A similar attempt is proposed in [5], where relevance is measured with Partial Conditional Entropy which allows to gradate it between 0 and 1 as well as to avoid calculation of the conditional probability density. For fuzzy performance queries [6] both text and spatial relevance dimension algorithm both grant to match geo-textual data with greater efficacy. For defining clinical relevance "post Hoc" analysis of HPV-negative search results is carried out on comparison basis by creating linear arrays [7]. Generating query refinement based on pseudo relevance feedback [8] is

found an efficient method for query modeling. HMS is applied for building predictive model performed by a recommender system [9]. Relevance-generating query refinement is adapted also for image retrieval in which case relevance score is generated before and after the user's feedback [10]. User-centered content is analyzed by applying interface design techniques in order to access the relevance of a German e-mental health information portal psychenet.de whose target audience is the group of patients with mental disorders undecided as to taking the treatment [11]. In Automatic Speech Recognition (ASR) spoken term detection relies mainly on the relevance feedback which improves the system accuracy from the perspective of the user need [12]. Learning the relevance from data rather than from the input vector can facilitate the retrieval process as well when considered is the aspect of nonlinear dimensionality reduction called relevance units latent variable model [13]. Relationships between query and document sentences grouped thematically are studied for mutual ranking-based both during and after relevance propagation on comparable basis [14] which indicates that the algorithm after relevance propagation is more efficient than during it.

In the proposed HCI scenario, while searching, the user gathers the information indexed by the standard databases and depending on the system's efficiency the process is continued by either removing the key phrases that are found non-relevant or by expanding the query making it more likely to match the information. Such a scheme enables us to analyze the dynamism of the precision elevation trend for the query types.

3 Generating Relevance Under HMS

Our approach is based on the real-time analysis while entering the query segments which are defined as the keywords, the phrases, the expressions or the sentences. The user has no sufficiently explicit knowledge about the subject or the field of the information which he really needs. For this reason, the queries are often profiled e.g. the system suggests the keywords or phrases with the highest possible relevance via the pull-down menu. As a result, submitting the first query segment generates a distracted result that prevents the user to check the contents of the individual links suggesting relevancy from his the viewpoint. The ranking list contains the results at the mutually distant positions, sometimes every few tens of links, and this causes understandable discouragement to continue the searching process.

Nevertheless, the stage of the first query segment submission is finished by acquiring a certain knowledge, whereby the information need becomes more clarified. In this case, the subsequent submission of the query generates a higher relevancy by profiling either the mode or keywords occurring with the higher frequency in the digital database. The process is shown in Fig. 1.

In the course of Interactive Gradually Generating Relevance QR under HMS, each query is divided into segments, depending on the dynamics of changes in the

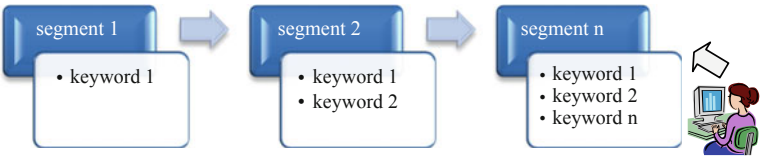


Fig. 1 Submitting queries with Interactive Gradually Generating Relevance QR under HMS

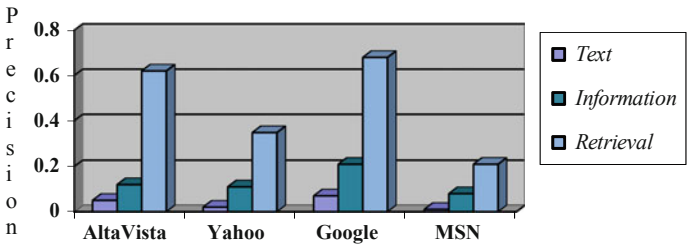


Fig. 2 Comparative table and the diagram of the search results obtained with Interactive Gradually Generating Relevance QR under HMS

ranking list of results. As shown, the user enters the first segment (here, a keyword) and analyzes the ranking list of the search results. The most frequently occurred keyword following the keyword just entered gives a rough idea of the type of search information. As the user continues the process, he or she comes to the conclusion that the length of the ranking list is becoming shorten allowing to click more of the search results, while those inadequate to his information needs are automatically eliminated making the relevant results moving towards the top position.

Figure 2 shows how to reduce the number of search results by adding, removing, or replacing more keywords. The dynamics given of the regular decrease of the search results corresponds to the dynamics of the growing relevance value that match the query. This means that while the user gets less and less results, the number of the results deemed by him to be relevant increases with the dynamics of an equivalent decline ago. The bar graph in Fig. 2 indicates the growth, presumably generated by the user working under HMS.

4 Metric of Generative Relevance Under HMS

The measure of the translingual systems' effectiveness is the interpolation rule [15]. It allows simultaneous correlation of the most popular measures, which are precision and recall. The interpolation rule graph consists in determining the nodes, which are usually defined at the standard levels of recall and measurement points which divide the interval {0,1} into the sections of equal length. The number of segments corresponds to the number of results deemed to be relevant. Following the

interpolation rule, at the standard level $l \in \{0,1\}$ for any level of recall, the precision value $p \leq l$ is maximized.

In addition to traditional approach like correlation of precision and recall, in the approach developed called Interactive Gradually Generating Relevance QR under HMS it is assumed that the system responses can be partly relevant, that is although the user finds them relevant, they seem not to meet his or her information need to the full, at least to click on the link of the particular document. Their role is limited to suggestion of a query refinement like for instance adding another or replacing existing query segment. Such an assumption is essential in the sense that although the user's need is still not fully met, he or she is somehow guided by the dynamism of changes in the ranking list of the system responses.

For the set Z of all the system responses, defined is relevance variable $p \in N$ of $\mathcal{R}(p)$—the complete relevance function, $P\mathcal{R}(p)$—partial relevance and $I\mathcal{R}(p)$—the function of the results assumed by the user to be irrelevant.

$$Z: = \{p \in N: \mathfrak{R}(p), P\mathfrak{R}(p), I\mathfrak{R}(p)\} \tag{1}$$

Additionally, we assume that for any relevance coefficient $\varphi \in \langle 0, 1\rangle$ there exists such a relevance variable $p \in N$ that it is true that

$$\bigwedge_{0 \leq \varphi \leq 1} \bigvee_{p \in N} \lim_{p \to \infty} \frac{\varphi \mathfrak{R}(p)}{\varphi[\mathfrak{R}(p) + P\mathfrak{R}(p)] + (1-\varphi) I\mathfrak{R}(p)} = 1 \tag{2}$$

Assuming the above criteria of the results' classification which is dependent upon the level of the user's satisfaction from meeting the information need, the measure of the relevance generated $\bar{P}_i$ which is the quotient of the sum of the products of $P_k(i)$ precision interpolation system for the each selected result multiplied by index $k \in \{k_1, k_2, k_3, \ldots, k_n\}$ to the number of results retrieved W_i. The index α of the user satisfaction level has a value between $\alpha \in [0, 1]$.

$$\bar{P}_i = \frac{\sum_{i=1}^{n} [P_k(i) \times k_i]}{W_i} \tag{3}$$

$$k_i = \begin{cases} 1 & for \quad \mathcal{R}(p) \\ \alpha_i & for \quad P\mathcal{R}(p) \\ 0 & for \quad IR(p) \end{cases} \tag{4}$$

For the example given above, the value of the index α of the user satisfaction level at $\alpha \in \{1, 0.7, 0.5, 0.3\}$ of the information need at the recall levels $P_k \in \{0.25, 0.5, 0.75, 1\}$, respectively. According to formula

$$P_k = \alpha_k \frac{R(k)}{W(k)} \tag{5}$$

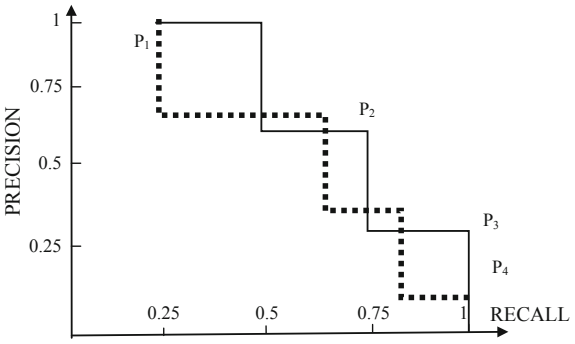


Fig. 3 Comparative precision-recall graph of the *solid line* and *dotted line* plot of interpolation rule and Interactive Gradually Generating Relevance QR under HMS

Table 1 Contingency table of interpolated recall-precision metrics

Interpolation rule	TREC metric	Interactive Generating Relevance QR under HMS
Recall	0.27	0.14
Precision	0.20	0.10

wherein $R(k)$ is the number of the relevant results, and $W(k)$ is the total number of the responses returned, the interpolated precision nodes P_k are the points on the X-axis having the following coordinates $P_1(0.25, 1)$, $P_2(0.5, 0.7)$, $P_3(0.75, 0.37)$ and $P_4(1, 0.08)$.

Figure 3 presents a comparative precision-recall graph of interpolation indicated by the solid line while following the rule of dividing the results into relevant and non-relevant and using the method for assessing the effectiveness of the search process, in which interpolation takes into account α—the level of the user interest in the document. The interpolation chart of precision graded by the user's interest is marked by a dotted line. The value of the level is a value of likelihood of adding another segment to the user's query with the aim at orienting his knowledge in accordance with the demand for information. Subsequent iterations move on the user closer towards the stage when he realizes that the query is correlated with the possibilities of information retrieval by the system to the extent that he can acknowledge his information need is fully satisfied. Such a stage is reflected by systematic decrease in the total number of search results in favor of moving the relevant results towards the top positions of the ranking list. This way a set of relevant results gradually maps the set of all the retrieved results. The precision values interpolated by the level of the user's interest gradually goes to 1.

As shown in Table 1 our measure of recall-precision interpolation generated by the user has a value significantly lower than the one compared with the measure used for many years by the American Text Retrieval (TREC) Conference series organized by National Institute of Standards and Technology (NIST),[1] which is

[1]Text Retrieval (TREC) conference series http://trec.nist.gov/.

binary relevancy. This is due to conditioning of the relevance value on the probability of adding another segment to the query and the associated with it the higher accuracy of measuring the search efficiency. In translingual systems, where efficiency is additionally conditioned upon the translation quality of the query, this seems of particular importance.

5 Bilingual Ad-Hoc Framework

In order to provide a detailed information on the experiment reported, this section introduces an example of how our method works as well as presents a relevance-oriented bilingual Ad Hoc framework. Prepared is a set of English Ad Hoc tasks [16] for which we profile our queries by gradually refining them under HCS. For each Ad Hoc task, we define the relevance criteria signifying a clear distinction between relevant and non-relevant criteria as presented in Fig. 4 below.

Our evaluation is based on the system reaction to the Ad Hoc tasks which apart from providing the overall idea of the information content, specifies the difference between the relevance and irrelevance in order to extract only the snippets of the text that semantically relate to the query and comprise with the target language syntax.

A group of thirty students of the Faculty of Computer Science and Management from Wroclaw University of Technology was entrusted with fifty one-up-to-three-word queries to be entered in the Ad Hoc mode. The queries were divided into segments one up to three being single words in a sequence. Each student received five Ad Hoc tasks to enter them sequentially. Relevancy was generated based on an analysis of the adequacy of the information in the documents retrieved to their definition of relevance explicitly specified in the tag <PL-Narr>.

The rating presumably depends on the user's subjective conviction about the level of relevance contained in the interval [0, 1]. Subjectivity depends in fact, on several factors including foreign language skills, e.g. for the user the document

```
UNESCO World Heritage Sites

Give the names and/or location of places that have been
designated as UNESCO World Heritage Sites of outstanding
beauty, or importance.

Relevant documents must mention the name and/or
geographical location of monuments, cities or places that
have been officially designated by UNESCO as World
Heritage sites of outstanding beauty or importance.
Discussions on potential or candidate sites are not
relevant. It must be clear from the document that the
official UNESCO status is involved.
```

Fig. 4 A sample Ad Hoc task from the CLEF set [16]

language is unknown, the evaluation score is 0, because he is not capable to read the information. The research tool were the search engines. It should be emphasized that the students were not familiar with the directly proportional dependence between the sequential addition of the segment to the query, and the dynamics of changes in the search precision value. Thus, the results of the experiment were deprived of a possible subconscious intention of subordination of the subjective assessment of the user to the analyzed dynamics of this relationship growth.

Each student submits the queries one by one to the selected by us distributed search systems starting from the one-word query to the full form given in the test prepared by us as presented in the task—in this case "UNESCO World Heritage Sites" that starts from "UNESCO", then "UNESCO World Heritage" following the system prompt to be refined to the full form being "UNESCO World Heritage Sites". Every system reaction is evaluated based on the relevance of the first 20 results. The Assessor acting as the End-User, selects manually the documents depending on the target language, the entity, the document content agreement with the task. We register the number of the system overall responses, the number of the relevant documents and the query class. Any changes in proportion between the relevant to total responses for each query class are reported. For every other search strategy the same procedure is carried out. Finally, we repeat the process with another search system.

In addition, we consider the search modes so as to compare the resulting lists of documents in French. Our starting point is that a user builds his knowledge based on the system responses.

The research question posed is then how a user can make the most out of the search systems and which strategies prove the most efficient in this kind of interaction HCS. So we attempt to specify the classification of the multilingual information systems, analyze the changes in proportion between the relevant and non-relevant search results in multilingual settings, study a dynamism of the system precision elevation trend based on different language models and finally, we compare the translingual search strategies, which include the search modes, completion, deletion, or replacement the keywords one by another.

6 Dynamism of Precision Elevation Trend

The increasing trend is dependent upon the relevance of the $n - 1$ step in terms of its overall influence on the precision. A query segment, either a phrase or a word that produces worse results than the one recorded in the previous step is replaced with a new one that can be its synonym more commonly used in the net. Therefore, the relevance is layered one upon another whereas the set of the system responses is changing its contents from the original which is the one resulting usually from a one-word query to the one that consists of the gradually increasing number of the relevant results. Under the optimum conditions, the resulting list consists of the

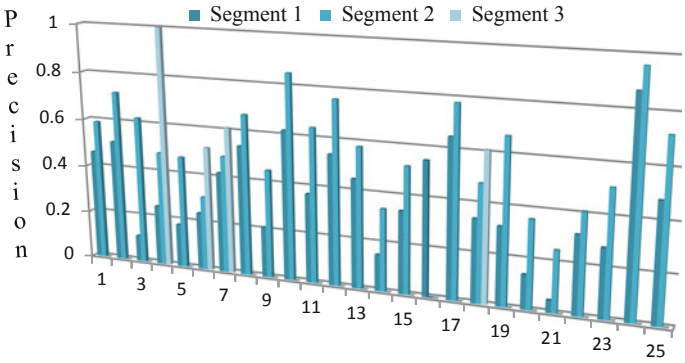


Fig. 5 The diagram of the precision elevation trend when Interactive Gradually Generating Relevance QR under HMS is applied for the first 25 Ad Hoc tasks

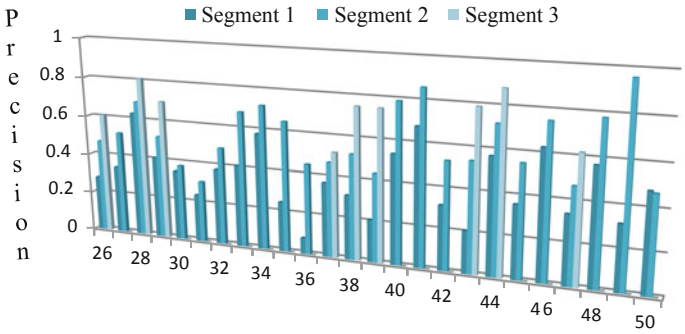


Fig. 6 The diagram of the precision elevation trend when Interactive Gradually Generating Relevance QR under HMS is applied for the last 25 Ad Hoc tasks

relevant responses only, short enough for the user to meet the information need in full.

In Fig. 5 noticeably the greatest dynamism of the precision growth is observed for query 3, in which case the difference between the first and last query segment is maximal when compared to the other queries. On the other hand, we notice the results of higher precision level for which the growth value rate is rather low like question 24.

In the second part (Fig. 6) there are many more 3-segment queries what enables more accurate analysis to determine the rate of precision dynamics growth of the search. The last result is an exception, because adding a segment did not cause the increase in the value of precision. The highest growth in this group went to query 49 and 36 at a significantly lower overall result.

Our methodology indicates that narrowing the ranking list to the relevant results only is an effect of the user's systematically increasing knowledge about the topic

for which the information is searched. In other words, at the beginning the user's knowledge is quite general and so is the query. Based on the information gained from the digital documents retrieved, the user is becoming more and more aware of the specific information in need. Eventually, the query consists of the phrases or words precise enough to fulfil completely the information need.

7 Conclusion

In contrast to existing approaches, we develop a novel user-centred methodology that significantly influences the information system efficacy. Due to the number of data, in particular, the same queries submitted to three engines, each of which being refined systematically, we focus on the methodology rather than on presenting all the diagrams with the research results in detail. However, the overall results are presented and discussed.

The starting point of our research is that by broadening his or her knowledge about a particular issue, a user is becoming more aware of the information need which is transformed by entering the more precise queries. We noticed that the novel technique proposed generates significant changes in proportion between the relevant to the total system responses. Depending on the databases accessible by the system and the query profile, presumably sometimes also some other factors like the system self-scoring algorithm, or matching algorithm, it is possible to achieve a ranking list of almost all the relevant results with the defined relevance criteria for each search. One of the observations made is that it is noticeable that what is called in literature Invisible, Hidden or Deep Web is in fact not quite precisely defined—no one knows the document types, URLs, or Web sites and especially the search engines that can be classified as definitely those that crawl the net inaccessible for other systems.

In the further stage, we plan to concentrate on the precision trend in order to develop a methodology for intensifying the set of the relevant results even more for some other language pairs like Polish, English and French.

References

1. Mizera-Pietraszko, J., Zgrzywa, A.: Vertical search strategies in federated search environment. In: Zgrzywa, A., Choros, K. (eds.) Advances in Intelligent and Soft Computing, pp. 171–180. Springer (2010)
2. Mizera-Pietraszko, J.: Multilingual document mining for unstructured information. In: Pahikkala, Vayrynen, Kortela, Airola (eds.) Proceedings of the 14th Finnish Artificial Intelligence Conference, Aalto University School of Science and Technology, Publications of the Finnish Artificial Intelligence Society 25, Espoo, Finland (2010)
3. Mizera-Pietraszko, J.: Interactive Document Retrieval from Multilingual Digital Repositories, pp. 423–428. IEEE Xplore Digital Library, IEEE Computer Society (2009)

4. Mizera-Pietraszko, J.: Method for evaluating impact of machine translation quality on multilingual information retrieval. Ph.D. Dissertation, Wroclaw University of Technology (2014)
5. Zhong, M., Liu, L., Lu, R.: A new method of relevance measure and its applications. In: Proceedings of Advanced Information Processing and Web Technology (ALPIT 2007), pp. 595–600 (2007)
6. Li, J., Wang, H., et al.: ILP heuristics and a new exact method for bi-objective 0/1 ILPs: application to FTTx-network design. Geoinformatica **20**(3), 453–469 (2016)
7. Petry, K.U., Cox, J.T., Johnson, K.: Evaluation HPV-negative CIN2+ in the ANTENA trial. Int. J. Cancer **138**(12), 2932–2939 (2016)
8. Wang, J.: The study of methods for language model-based positive and negative relevance feedback in information retrieval. In: Proceedings on International Symposium on Information Science and Engineering (ISISE 2012), pp. 39–43. IEEE Computer Society (2012)
9. Akuma, S., Iqbal, R., Jayne, Ch., et al.: Comparative analysis of relevance feedback methods based on two user studies. Comput. Hum. Behav. J. 138–146 (2016)
10. Kumar, K.K., Gopal, T.V.: Multilevel and multiple approaches for feature reweighting to reduce semantic gap using relevance feedback. In.: Proceedings of 14th International Conference on Contemporary Computing and Informatics. IEEE Computer Society (2014)
11. Dirmaier, J., Liebherz, S., Saenger, S.: Psychenet.de: development and a process evaluation of an E-mental portal. Inf. Health Soc. Care **41**(3), 267–285 (2016)
12. Lee, H., Chen, Ch., Lee, L.: Integrating recognition and retrieval with relevance feedback for spoken term detection. IEEE Trans. Spoken Speech Lang. Process. **20**(7), 2095–2110 (2012)
13. Gao, J., Zhang, J., Tien, D.: Relevance units latent variable model and nonlinear dimensionality reduction. IEEE Trans. Neural Netw. **21**(1), 123–135 (2010)
14. Cai, X., Li, Wu: Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization. IEEE Trans. Audio Speech Lang. Process. **20** (5), 1597–1607 (2012)
15. Voorhees, E.M., Lori, P., Buckland, T.: Common evaluation measures, NIST Special Publication 500–274. In: The Proceedings of the 16th TREC Conference, Gaithersburg, Maryland, 5–9 Nov (2007)
16. Mizera-Pietraszko, J.: Accuracy of the AID system's information retrieval in processing huge data collections. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) CLEF 2006 Workshop. Working Notes Alicante, Spain, 20–22 Sept 2006. GEIE-ERCIM, Sophia Antepolis, 7, ERCIM, Italy (2006)

Definition of Requirements for Accessing Multilingual Information and Opinions

Jan Derkacz, Mikołaj Leszczuk, Michał Grega, Arian Koźbiał and Kamel Smaïli

Abstract With the development of the Internet and satellite television, access to thousands of programs and messages in different languages became widespread. Unfortunately, even well educated people do not speak sufficiently in more than two or three foreign languages, while most know only one, and this significantly limits the access to this information. In this paper, we define requirements for an automated system for Accessing Multilingual Information and opinionS (AMIS) that will help in the understanding of multimedia content transmitted in different languages, with simultaneous comparison to counterparts in their native language user. The concept of understanding we use will provide access to any information, regardless of the language in which it is presented. We believe that the AMIS project can have a immense and positive impact on the integration and awareness of society in social and cultural terms.

Keywords Multilingual translation · Automated translation · Opinion mining · Sentiment mining · User requirements survey · Architecture design · AMIS · Video summarization

J. Derkacz · M. Leszczuk (✉) · M. Grega · A. Koźbiał
AGH University of Science and Technology, Kraków, Poland
e-mail: leszczuk@agh.edu.pl

J. Derkacz
e-mail: derkacz@kt.agh.edu.pl

M. Grega
e-mail: grega@kt.agh.edu.pl

A. Koźbiał
e-mail: ariankozbial@gmail.com

K. Smaïli
University of Lorraine, Nancy, France
e-mail: Kamel.Smaili@loria.fr

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_24

1 Introduction

With the development of the Internet and satellite television, access to thousands of programs and messages in different languages became widespread. Unfortunately, even deeply educated people do not speak sufficiently more than two or three foreign languages, while most know only one, and this significantly limits the access to this information.

Therefore, there is a new challenge: how a user can access the information and opinions in foreign languages? Due to cultural and political differences, it is very often necessary to compare the information in its mother tongue to the information provided on the same subject in a foreign language. For example: how AIDS is presented in Saudi Arabia, and the like in the US? What is the opinion of "The Jerusalem Post" on Yasser Arafat? And what, for example, "Al-Quds" (Palestinian newspaper: http://www.alquds.co.uk/) gives about the same topic?

In this paper, we define requirements for an automated system for Accessing Multilingual Information and opinionS (AMIS) that will help in the understanding of multimedia content transmitted in different languages, with simultaneous comparison to counterparts in their native language user. The concept of understanding we use will provide access to any information, regardless of the language in which it is presented. The proposed system could eventually be integrated with remote-enabled TV, as well as any Web browser.

We define the process of understanding as the assimilation of the main ideas carried by recorded video. The best way to help the understanding is summarizing information. Therefore, AMIS will focus on the most relevant information, summarizing it, and presenting to the user. Another aspect of the AMIS system is to compare the two summaries from two languages, describing the same subject in a visual, audio or text way, as well as the presentation of the differences between their content in terms of information and opinions. Moreover, the system will obtain information from the Internet and social media that will be used to strengthen or weaken the received opinion. In summary, the following research methods will be applied in the AMIS project:

- Summarizing text, sound and video sequences. The proposed video summarization will support cross-lingual opinion studies and will be based on high-speed detection of shots in video content for abstracting digital video [4]. The proposed text extraction approach would be based on video segmentation [8].
- Automatic Speech Recognition (ASR)—what is still a difficult task [2] because of disfluencies, ill-formed sentences, and increased speaking rate. In the previous research, the diachronic aspect was tackled, and research relied on text documents of the same period to find new relevant words and to improve the ASR performance [1]. In AMIS, where the goal is to process video documents of the day, other, more innovative approaches will have to chosen.
- Machine translation is used to translate the output of the speech recognition into the target language. Translation will be developed for the following language pairs: French-English, French-Arabic, and English-Arabic. The system will use the state-

of-the art Moses system along with the Giza++ toolkit [3], as well as the SRILM language modelling toolkit [7], as proposed in the Moses' scripts.
- Interlingual analysis of the opinions and feelings. This is a challenging research area for which an original automatic method is proposed, allowing to tag a text in terms of opinions by transferring the annotations from the domain of movie reviews to other domains (news and talks) [5, 6].
- The achievement of a successful synergy of previous research topics.

The remainder of this paper is structured as follows. User requirements will be defined at first in Sect. 2. Based on this functional and architectural requirements, recommendations will be provided in Sect. 3. Finally, Sect. 4 summarizes an impact of expected results.

2 User and Services Requirements

In order to know potential end-users' expectations and opinions a questionnaire was prepared. The idea was to provide a possibly concise and clear form for the respondents that will be at the same time informative for research teams.

The questionnaire was organized in the following main parts: Respondent information, Usage of sources of information, Functionalities of the proposed system and Human interface. The questionnaire was made available at a web page elaborated by one of the project partners. Until now over 80 individuals have given their answers. In most questions respondents were asked to choose from a number of options, to make a ranking of certain features or to give a specific number (such as average number of watching news per day). Finally, persons answering the survey could give their opinions and recommendations as a free text.

Summary of the preliminary results of the survey is provided below.

2.1 Respondent Information

In the first part of questionnaire information concerning the respondent was provided. This included information about gender, age and mother tongue.

2.1.1 Gender and Age

The number of men (66 %) almost doubled the number of female among the respondents.

Young and medium age persons prevailed among persons answering the questionnaire (Fig. 1).

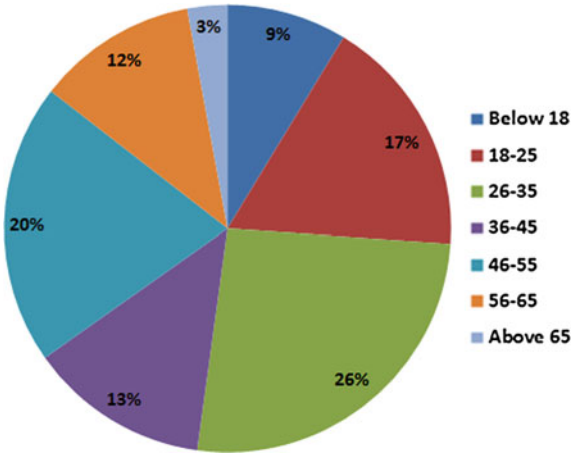


Fig. 1 Age

2.1.2 Mother Tongue

Most of persons answering the questionnaire speak French, Polish and Spanish. This results from the project consortium composition. Web page with the questionnaire was public however most of the respondents were requested by the project teams among persons they know.

Category "other" includes also cases where no answer was given—which means that this can also contain persons with mother tongues mentioned above (Fig. 2).

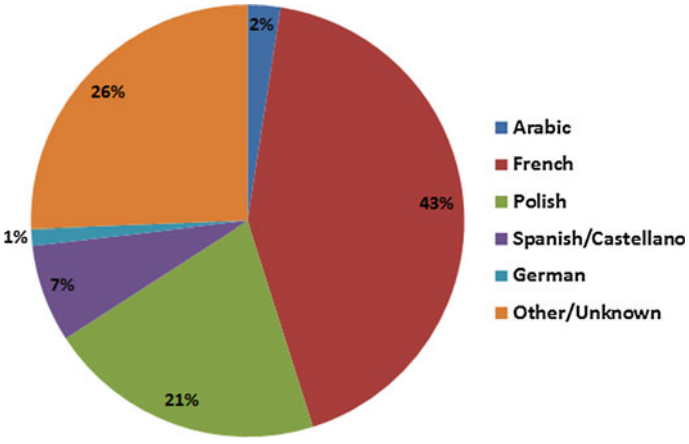


Fig. 2 Mother tongue

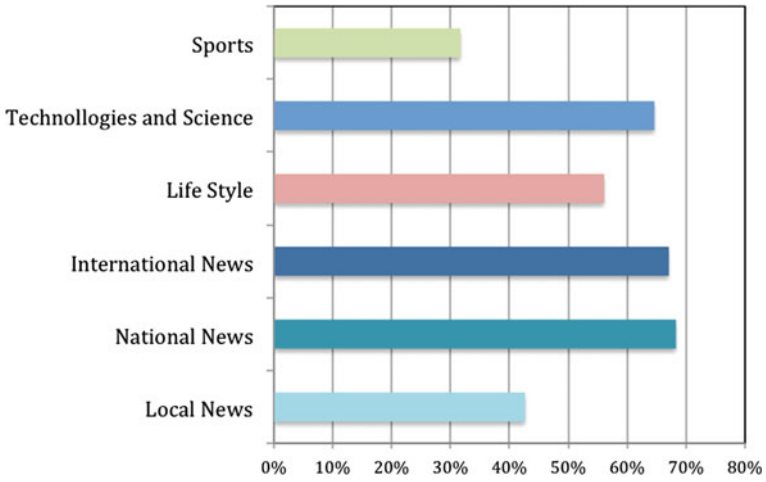


Fig. 3 Interest in subjects

2.2 Usage of Sources of Information

The second part of the questionnaire was dedicated to collecting information related to users preferences concerning thematic subjects s/he is interested in. Subjects were organised into 4 main categories: Sports, Technologies and Science, Life Style and News (International, National, Local).

2.2.1 Interest in Subjects

The highest percentage of persons are interested in International and National News. Relatively high number of respondents are interested in the category "Technologies and Science". This can be biased by the fact how the respondents were chosen (Fig. 3).

2.2.2 Preferred Forms of Programs

Respondents had a choice of three main programs types: discussion of experts, news supported with or films presenting a given subject. The last two categories have almost equal number of preferences. Discussion of experts is the least popular among the three (Fig. 4).

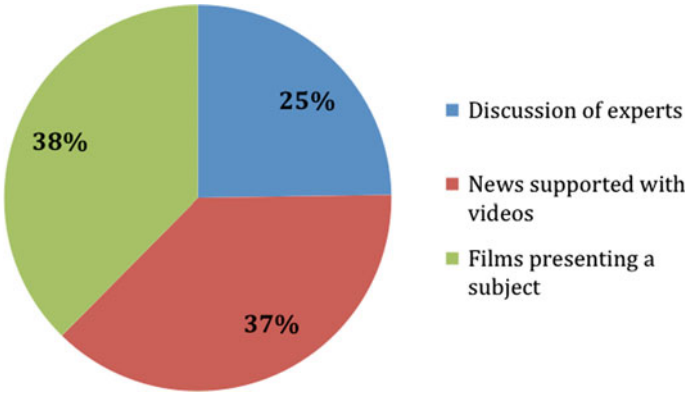


Fig. 4 Preferred forms of programs

2.2.3 Usage of News on TV and Internet

The average frequency of TV news watching is on the level of 0.6—it appears that according to given feedback, significant majority of respondents do not watch TV news. Those who watch news on TV usually spend on it several minutes per day.

Relevant numbers related to Internet news are different: news are browsed a few times per day on average, and average watching session is by 50 % longer that this on TV.

2.3 Functionalities and Human Interface of Proposed System

Finally, opinions on functionality and human interface were collected. Feedback was received on such features as: preferred translation, type of equipment which would be used by end-user.

2.3.1 Preferred Translation

Large majority of the respondents would prefer to have the translation from a foreign language to be done in a form of a text displayed over the original content (Fig. 5).

2.3.2 Equipment Preferred by End-Users

Respondents were given a question: "which equipment would you prefer to use with the proposed functionalities". Almost 70 % have indicated computer. Least fre-

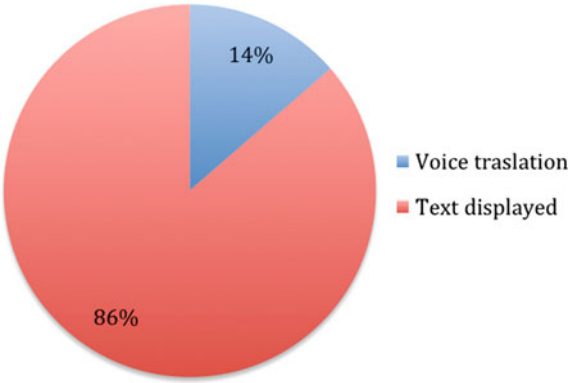


Fig. 5 Preferred translation

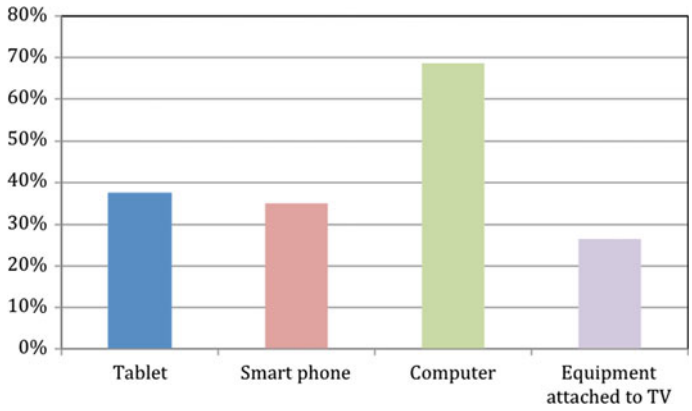


Fig. 6 Equipment preferred by end-users

quently chosen was equipment attached to TV set. In the free text comments Xbox One was indicated once (Fig. 6).

2.4 Selected Suggestions Given by Respondents

For internet services such as Pinterest, for which there are lot of activities, tutorials and resources in different languages, "Google Translate" can be used.

It was recommended that end-user interface should have a functionality allowing for copy-pasting a video link and then choosing between a shorter video, an oral or written translation or further information. The rank of the functionalities depends on what sort of information a user wants to find.

It was pointed out that for end-user it would be interesting not only to compare information given by different TV stations in a country but also to show how news on a given subject is presented in different countries.

Additional information could be displayed over the video for instance as some additional pictures about the topic presented in the video at this moment.

2.5 Summary

The survey might be broadened in different respondent categories in order to be more representative. The first results have also shown that additional questions might be useful for having a clearer and more precise view of end-users' expectations and requirements.

3 Overall Architectural Recommendations

During the initial phase of the AMIS project we have proposed a two phase approach to the software architecture. We were working towards a solution that would, on one hand, result in an effective system and, on the other hand, allow for seamless integration of components delivered by the Project Partners.

In the first phase we have created a simple pipeline architecture. This architecture will be at the same time effective in delivering summarized and translated content and will be fairly easy to implement and integrate. This approach will allow to analyse future research and engineering challenges in order to adapt the second phase architecture. We have also agreed that in the first phase the system will be lacking some of the planned non-critical features such as full social networks integration and sentiment analysis. Proposed phase one architecture is depicted in Fig. 7 with the main pipeline denoted.

We start with a full video file which is summarised to desired length. The audio of the summarized video file is extracted and passed to the speech recogniser. The result of the speech recognition are translated using the machine translation algorithms. The summarised video file with added captions provided by the machine translation is presented to the user. At the same time both video summarization,

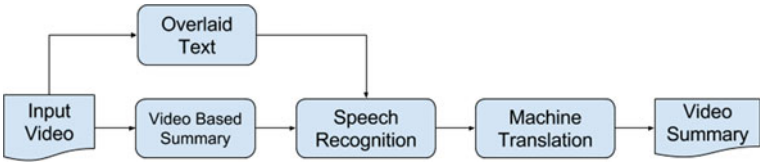


Fig. 7 First phase architecture

speech recognition and machine translation are supported by information provided by the optical character recognition system that is analysing the overlaid video text.

The second phase is yet to be defined, but will incorporate the missing features—namely integration with social media and sentiment analysis. When moving to the second phase we also plan to enhance functionalities and algorithms used in the first phase in order to deliver higher quality summarizations and translations.

4 Summary on Impact of Expected Results

We believe that the AMIS project can have a immense and positive impact on the integration and awareness of society in social and cultural terms.

First of all, AMIS will greatly minimize difficulties in the exchange of information regardless of language differences, which nowadays is extremely important.

Its use concerns mainly the domain of social and cultural, as it allows one to explore another side of the event. This will allow users to understand the comments, as well as indirectly the culture of other nations.

At the scientific level the project concerns synergies and new research linking summarizing video, audio and text materials, ASR, machine translation and opinion exploration.

Acknowledgments Research work funded by the National Science Centre, Poland, conferred on the basis of the decision number DEC-2015/16/Z/ST7/00559.

References

1. Illina, I., Fohr, D., Linares, G.: Proper name retrieval from diachronic documents for automatic speech transcription using lexical and temporal context. In: Workshop on Speech, Language and Audio in Multimedia, Sept 2014. Penang, Malaysia. https://hal.inria.fr/hal-01092224
2. Jousse, V., Linares, G.: Spontaneous speech characterization and detection in large audio database. In: SPECOM (2009)
3. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Demonstation Session, pp. 177–180 (2007)
4. Leszczuk, M., Papir, Z.: Accuracy vs. speed trade-off in detecting of shots in video content for abstracting digital video libraries. In: Protocols and Systems for Interactive Distributed Multimedia: Joint International Workshops on Interactive Distributed Multimedia Systems and Protocols for Multimedia Systems, IDMS/PROMS 2002 Coimbra, Portugal, 26–29 Nov 2002. Proceedings, pp. 176–189. Springer, Berlin, Heidelberg (2002). http://dx.doi.org/10.1007/3-540-36166-9_16
5. Saad, M., Langlois, D., Smaïli, K.: Comparing multilingual comparable articles based on opinions. In: Proceedings of the 6th Workshop on Building and Using Comparable Corpora. pp. 105–111. Association for Computational Linguistics ACL, Sofia, Bulgaria, Aug 2013. https://hal.inria.fr/hal-00851959

6. Saad, M., Langlois, D., Smaili, K.: Building and modelling multilingual subjective corpora. In: Chair, N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland, May 2014
7. Stolcke, A.: Srilm—an extensible language modeling toolkit. In: ICSLP, pp. 901–904. Denver, USA (2002)
8. Weber, J., Lefèvre, S., Gançarski, P.: Spatio-temporal quasi-flat zones for morphological video segmentation. In: Mathematical Morphology and Its Applications to Image and Signal Processing: 10th International Symposium, ISMM 2011, Verbania-Intra, Italy, 6–8 July 2011. Proceedings, pp. 178–189. Springer, Berlin, Heidelberg (2011). http://dx.doi.org/10.1007/978-3-642-21569-8_16

Query Answering to IQ Test Questions Using Word Embedding

Michał Frąckowiak, Jakub Dutkiewicz, Czesław Jędrzejek, Marek Retinger and Paweł Werda

Abstract This paper presents an improvement over the results of Wang et al. on query answering to IQ test questions using word embedding. The improvement comes from using the Glove method and renormalization of results using the best results of two leading methods. This latter approach combines knowledge coming from different corpuses.

Keywords Query answering · IQ testing · Word embedding · Glove method

1 Introduction

Query answering is one of the cognitive processes, where the dominance of humans decreases rapidly in favor of machines in the last years. Changes to the status quo are influenced mainly by the increased computation capabilities of the modern hardware, enabling analysis of large corpora and constantly improved word embedding methods. A path to significant progress came from The Text REtrieval Conference (TREC) Query Answering track, last in 2007 [7] fostering extensive research on different question types and finding answers in different kinds of corpora. Later The Text Analysis Conference (TAC) continued the TAC QA track [24] which evolved into the Knowledge Base Population (KBP) track, where the task was factoid-question answering.

The greatest advantage of approaches proposed at TREC and TAC was the preparation of large test collections that could serve as standardized accessible corpora, and a cross evaluation of methods.

M. Frąckowiak · J. Dutkiewicz · C. Jędrzejek (✉) · M. Retinger · P. Werda
Institute of Control and Information Engineering, Poznan University of Technology, Poznań, Poland
e-mail: czeslaw.jedrzejek@put.poznan.pl

J. Dutkiewicz
e-mail: jakub.dutkiewicz@put.poznan.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_25

The next enormous step in NLP related artificial intelligence was IBM Watson [13]. Watson is able to handle complex sentence questions of 13 classes and is largely domain-independent. One of those classes is multiple-choice question. A question analysis is built on a foundation of general-purpose parsing and semantic analysis components. The Question Classes evaluation in 2012 over all the question classes, showed an F measure of 0.637 [5]. For the multiple-choice question class a value for precision is 0.650, and a value for recall is 0.684 giving $F_1 = 0.667$.

One of the areas where machine Query Answering results could be of significant importance is Intelligence Quotient Test (IQ). Since IQ test scores are considered as a measure of intelligence, studying the area could lead to improved comparison of human and machine intelligence [1, 9, 12].

Common IQ tests contain several categories of questions: verbal comprehensive questions, which usually constitute a large proportion (near 40 %) of questions, mathematical questions, logic questions, as well as picture questions [6, 10]. In this paper we are focusing only on verbal questions. The goal is to verify recent results of Microsoft Research China and University of Science and Technology of China [26].

The starting point is the use of deep learning for natural language processing (NLP), namely word embedding methods, that are easier to apply than semantic technologies using parsers [11]. Word embedding methods have been used in multiple research projects that focus on the ability of machines to understand the meaning of words and their correlations, rather than relations. The results obtained in [26] created a lot of publicity under a controversial title "Deep learning machine beats humans in IQ test".

The authors of [26] claim that a straightforward application of word embedding could not give satisfactory results, because standard word embedding technologies learn one embedding vector for each word based on the co-occurrence information in a text corpus. Since verbal comprehension questions in IQ tests usually consider the multiple senses of a word one has to modify approach of Mikolov [11].

We carefully inspect the results of [26] and extend it using the Glove method combined with heuristic methods. We also aim for finding other improvements.

Section 2 presents the scheme used in [26]. Section 3 describes our results using the Glove method and some heuristic, Sect. 4 analyses further potential improvements. Conclusions and future work are given in Sect. 5.

2 The Basic Word Embedding Approach

The starting point of the [26] analysis is the skip-gram method in word2vec [11]. Skip grams are word windows from which one word is excluded, an n-gram with gap. With skip-grams, given a window size of n words around a word w, word2vec predicts probability $p(c?w)$ of contextual words c. Given a sequence of training text stream $w_1, w_2, \ldots, w_K$, the objective of the skip-gram model is to maximize the following average log probability:

$$L = \frac{1}{K} \sum_{k=1}^{K} \sum_{-N \le j \le N, j \ne 0} \log p(w_{k+j} | w_k) \tag{1}$$

where w_k is the central word, w_{k+j} is a surrounding word, K denotes the length of the training text stream, and N indicates the context window which size is $2N + 1$. The conditional probability $p(w_{k+j} w_k)$ in the basic approach is expressed using the following *softmax* function:

$$p(w_{k+j} | w_k) = \frac{exp(v'^{T_{v w_k}}_{w_{k+j}})}{\sum_{w=1}^{V} exp(v'^{T_{v w_k}}_{w})} \tag{2}$$

where v_w and v_w are the input and output representation vectors of w during neural network training, and V is the vocabulary size.

Following [26] we use specific classifiers to recognize particular types of a verbal question (e.g., analogy I and II, classification, synonym, or antonym). The questions require verbal comprehension, i.e. recognition of a given question type. Solving for answers requires optimization of a function of word2vec pseudovectors, tailored to a given question type.

Particular questions have different kinds of relationships and the authors of [26] have to be credited with new solvers using distance criterion to answer multiple-type questions.

2.1 Analogy I

Analogy-I questions takes the form "*A is to B as C is to ?*". This task is very important because one can try to extend it to a logical representation of the notion of analogical proportion [22]. Showing that word2vec allows for solving this type a question led to enormous popularity of word2vec. The "*Man is to woman as king is to queen*" sentence became the leading phrase of popular artificial intelligence articles.

Originally, d is found by optimizing the following vector expression (v_k represent vectors). Choosing the distance measure is not straightforward. We will discuss this issue later. With cosine metrics the distance is

$$d = \underset{D \in T}{\text{argmax}} \cos(v_{(B)} - v_{(A)} + v_{(C)}, v_{D)}) \tag{3}$$

Wang et al. [26] proposed that instead of vectors for word as a string one should use separate vectors corresponding to all senses of a given word.

Wang et al. [26] argued that since Verbal questions in IQ tests usually consider the multiple senses of a word it is crucial to take this into account. They proposed a sophisticated method for relevant vector expression (not discussed here) and arrived at

$$D = \underset{i_a, i_b, i_c, i_{d\prime}; D\prime \in T}{\operatorname{argmax}} \cos(v_{(B,i_b)} - v_{(A,i_a)} + v_{(C,i_c)}, v_{(D\prime,i_{d\prime})}) \tag{4}$$

2.2 Analogy II

Analogy-II questions require two words to be identified from two given lists in order to form an analogical relation like "*A is to ? as C is to ?*". The best two candidate words B and D ate selected from two lists T_1 and T_2.

$$d = \underset{B \in T_1, D \in T_2}{\operatorname{argmax}} \cos(v_{(B)} - v_{(A)} + v_{(C)}, v_{(D)}) \tag{5}$$

2.3 Classification

The classification questions require a word to be identified from a given list of words T, whose meaning is different from the others. To calculate the most probable answer for the classification type of questions, we need to measure the center c of the answer tuple T of all word vectors w_i.

$$c = \frac{1}{|T|} \sum_{w_i \in T} w_i \tag{6}$$

We then compare the distance of all words within the tuple to the center of the group. The most distant word is chosen as the most probable answer w for the question.

$$w = \left\{ w_i \in T \,:\, d(w_i, c) = \max_{w_j \in T} d(w_j, c) \right\} \tag{7}$$

2.4 Synonyms and Antonyms

For the Synonym and Antonym types of questions we measure the distance between the given question word w_q and each word in the answer tuple T. As an answer w for the question we choose the word, which is the closest to the given word.

$$w = \left\{ w_i \in T \,:\, d(w_i, w_q) = \min_{w_j \in T} d(w_j, w_q) \right\} \tag{8}$$

The argument behind using the same formula (8) also for antonyms, is that antonym and the original word have similar co-occurrence information, based on which the embedding vectors were trained.

3 Main Results

We used the same set of 232 questions as in [26] in order to compare our results with the results of this work (Wang et al.). The main results are presented in Table 1.

3.1 Training Sets

At the beginning the system has to recognize a type of a question. Wang et al. [26] manually collected and labeled 30 verbal questions from the online IQ test Websites [27] for each of the five types and trained an one-vs-rest SVM classifier for each type. The total accuracy on the training set itself was 95.0 %.

In our work we used 129 questions in total as a training set out of [23] with the following number of training examples in particular categories (i.e. Analogy-I, 20; Analogy-II, 20; Classification, 30; Synonym, 30; and Antonym, 29). Data was converted into structured format (JSON), dividing each question to a triplet (question, answers, correct answers). Questions from the training set were vectorized using TF-IDF transform as weighting function for features. Such an input was used in training the one-vs-rest linear SVM classifiers. Total accuracy on a training set itself was 100 %. For a question type recognition we obtain slightly lower precision of 91.8 %. The only misclassified questions were 19 Antonyms questions, due to high similarity of question structure to the Synonym type. This has no effect on results since we did not use an relational offset vector for Antonym distance. As our test set we used the set published in [26].

3.2 The Results

Our main results are presented in Table 1.

The methods and corpora presented in Table 1 are the following:

1. **Latent Dirichlet Allocation Model (LDA)**. This was used in [26] as one of baseline models using standard classical distributional word representations, i.e. Latent Dirichlet Allocation (LDA) [4]. They trained word representations using LDA on wiki2014 with the topic number 1000.

Table 1 Comparison of results in [26] with our results in percentage of correct answers

The corpus and the method	Analogy-I	Analogy-II	Classification	Synonym	Antonym	**Overall**
Skip-Gram Model						
Wiki (Publication)	44.00	27.59	26.42	33.33	32.65	33.19
Google News	48.57	23.07	36.58	62.50	36.95	43.36
Glove Small	51.06	24.13	40.38	62.74	32.65	43.85
Glove Large	53.33	27.58	49.01	66.66	36.73	48.44
Glove Wiki	46.00	24.14	33.96	60.78	36.73	41.81
SG-1	38.00	24.14	37.74	45.10	40.82	38.36
SG-2	38.00	20.69	39.62	47.06	44.90	39.66
Multi-Sense Model						
MS-1	36.36	19.05	41.30	50.00	36.59	38.67
MS-2	40.00	20.69	41.51	49.02	40.82	40.09
Other						
LDA	28.00	13.79	39.62	27.45	30.61	29.31
HP Doctorate degree or candidate	55.33	37.93	58.49	71.77	70.69	58.84
RK	48.00	34.48	52.83	60.78	51.02	50.86

2. **Human Performance (HP)**. Here answers were provided by humans through Amazon Mechanical Turk [26]. In Table 1 we show result for the most advanced participants.
3. **Skip-Gram Model (SG)**. In this baseline, the authors of [26] applied the word embedding trained by skip-gram [11] (denoted by SG-1). The skip-gram to learn the embedding on wiki2014 was run with the following parameter set: the window size as 5, embedding dimension as 500, the negative sampling count as 3, and the epoch number as 3. The second option was using a pre-trained word embedding by Google [28] with the dimension of 300 (denoted by SG-2).
4. **Multi-Sense Model (MS)**. In this baseline, the authors of [26] applied the multi-sense word embedding models proposed in [2, 18] (denoted by MS-1 and MS-2 respectively). For MS-1, the authors of [26] used the published multi-sense word embedding vectors [19] by the authors of [18] in which they set 10 senses for the top 5 % most frequent words. For MS-2, the authors of [26] adopted the same parameters as MS-1.
5. **Relation Knowledge Powered Model (RK)**. This method the authors of [26] goes beyond word embedding done on wiki2014. The parameters were as for SG-1. For relational part the online Longman Dictionary was used as the dictionary in multi-sense clustering together with a public relation knowledge set, WordRep [3], for relation training.

The above results used in [26] are presented using grey background. Our results employed the following parameters:

6. **Wiki (Publication)**. The configuration most close to SG-1. The epoch number 5 instead of 3 as in SG.
7. **Google News**. The pre-trained model based on Google News [16]. Embedding vector size 300, the negative sampling count as 3.

8. **Glove Small**. Pre-trained model based on Glove approach [20] using common crawl data, accessible on [15]. Embedding vector size 300.
9. **Glove Large**. Pre-trained model based on Glove approach, with much larger word count 840 billions [14].
10. **Glove Wiki**. Model based on Glove approach, trained on English Wikipedia dump from 2014.

We used the same questions [25] as in [26].

3.3 Analysis of Results

Our results for Wiki (Publication) are overall worse than SG-1: better for Analogy-I, and Analogy-II questions but significantly worse for Classification, Synonym and Antonym types of questions. This can be partially explained by the fact that for Antonym we did not use relational offsets. This causes an incorrect answers if a synonym appears in an answer set T. The Glove method **gives much better results than obtained in** [26] in pure embedding category. The overall accuracy does not look impressive. In particular, the percentage of correct answers for Analogy-I is much lower compared to [20], where it exceeds 80 %; however, the questions we use are much more difficult. For simpler questions using supervised methods 100 % accuracy was claimed [8].

Overall, the Multi-Sense Model results in [26] show only a minor improvement compared to the SG results for question answering. Another possibility could be too a simplistic treatment of MS [26].

4 The Importance of Normalization

While inspecting the results for Glove Large and Glove Wiki, we observed that the top Glove Large vectors have magnitudes a few dozen percent larger than for Glove Wiki. This led us to investigate the dependency of results on normalization.

4.1 Normalization

We applied the L^2 norm to the two Glove data sets a data set constructed with 42 billion words with a vector size of 300 elements (glove.42B.300d) and a data set built with 840 billion words with a vector size of 300 elements (glove.840B.300d). We measured accuracy $acc(d)$ for the dataset d of achieving the proper answer to the questions and the impact $imp(d, d\prime)$ of the normalization process on the process of answering the question for the data set d and its normalized equivalent $d\prime$. We

have applied the calculation to the entire data sets using the following metrics. Let us assume that i_q is the proper answer for the q question, we measure the quality of the solver $a(q)$ for that particular question as

$$a(q,d) = \begin{cases} 1 & w_{q,d} = i_q \\ 0 & w_{q,d} \neq i_q \end{cases} \tag{9}$$

With the defined quality of the solver for certain questions we can define the *acc* for one dataset and *imp* for the dataset and its normalized equivalent with respect to the question type T.

$$acc(d,T) = \frac{\sum_{q \in Q_T} a(q,d)}{\overline{\overline{Q_T}}} \tag{10}$$

$$imp(d,d\prime,T) = \frac{|\sum_{q \in Q_T} |a(q,d) - a(q,d\prime)||}{\overline{\overline{Q_T}}} \tag{11}$$

As it is seen in Table 2, the normalization process has a relatively high impact on the *classification*, *synonymous* and *antonymous* question types while its impact is relatively low for both *analogy* question types. The normalization process has little to no impact on the accuracy for the *analogy* question types but we can observe a minor improvement in accuracy for *classification*, *synonymous* and *antonymous* question types. The normalization process results in a slight improvement in overall accuracy.

Table 2 Accuracy of choosing the correct answer and impact of the normalization across different question types

Data set	Type	$acc(d,T)$ (%)	$acc(d\prime,T)$ (%)	imp $(d,d\prime,T)$ (%)	$acc(d\prime,T)$ $-acc(d,T)$ (%)
glove.42B.300d	Classification	37.7	32.1	24.5	−5.6
glove.42B.300d	Analogy-I	50.0	52.0	4.0	2.0
glove.42B.300d	Analogy-II	20.6	24.1	3.4	3.5
glove.42B.300d	Synonyms	37.2	43.1	25.4	5.9
glove.42B.300d	Antonyms	55.1	61.2	14.2	5.1
glove.42B.300d	**All types**	**41.3**	**43.9**	**15.5**	**2.6**
glove.840B.300d	Classification	32.1	37.7	13.2	5.6
glove.840B.300d	Analogy-I	56.0	50.0	6.0	−6
glove.840B.300d	Analogy-II	27.5	31.0	3.4	3.5
glove.840B.300d	Synonyms	39.2	50.9	15.7	10.7
glove.840B.300d	Antonyms	59.1	65.3	10.2	6.2
glove.840B.300d	**All types**	**44.3**	**47.8**	**10.3**	**3.5**

Table 3 Missing words in particular corpora

Wiki (Publication) and GloveWiki	Google News		Glove Small	Glove Large
Irremissible	Isobar	Programme	Aoristic	Aoristic
	Pedicel	Rowel	Dentiform	Dentiform
	Aoristic	Defence	Dendriform	Dendriform
	Orison	Flavour	Pediform	Munsell
	Cygnus	Dishevelled	Irremissible	Martinmas
	Secant	Agonising		Pediform
	Realisation	Unequalled		Tarboosh
	Dentiform	Mumbojumbo		Irremissible
	Dendriform	Catalogue		
	Doughnut	Crabwise		
	Rowel	Pretence		
	Beaufort	Oriflamme		
	Munsell	Reynard		
	Auriculate	Indurate		
	Shoulderblade	Tarboosh		
	Harbour	Arquebus		
	Martinmas	Pentad		
	Candlemas	Irremissible		
	Heptagon	Mobilise		
	Pediform	Savour		
	Reniform			

4.2 Missing Words

Better understanding of the results is achieved upon inspecting the answers to individual questions. In particular, certain corpora have large number of missing words (Table 3). There are several methods for getting an answer in such cases. We ascribed the value of cosine 1 or vector 0 to a vector magnitude of a missing word in questions depending on a metrics related to a question type. Using this approach if a missing word occurs in a question, for our best results (Glove Large and Glove Wiki) the answer is almost always incorrect. If a missing word occurs in a multiple choice set there is a minor improvement.

Surprisingly, the number of missing words for Glove Large exceeds the number of missing words for Glove Wiki and Glove Small. The word *irremissible* appears once in Wikipedia (and is disregarded since the threshold was selected as 5 words) although the Google search finds 50000 documents for which such a word appears. This suggests that attaching few document from Google could improve results by 2 %.

Table 4 Correlation of correct answers between corpora

Corpora	1	2	3	4	5	6	7	8	9
Wiki (Publication)	0	1	0	0	0	1	0	0	0
Google News	0	0	0	1	1	0	0	0	0
Glove Small	1	0	0	1	1	0	0	0	0
Glove Large	1	1	0	1	1	0	0	0	0
Glove Wiki	0	1	0	1	0	1	0	1	0

4.3 Additional Results

It is interesting to look at measures of answers with regard to measures of words appearing in questions. The text of the question 11 (Analogy-I group) is the following. "*trireme is to ship as triptych is to:*"

Answers: spear, stand, pattern, panel, play.

Distances for answers are: spear: 0.0460, stand: 0.2894, pattern: 0.1641, panel: 0.317, play: 0.164. The chosen answer: panel is the correct answer. However, in this example the correct answer (like in many cases) is not even in the 10 largest measures of closeness to the question vector: [(u'painting', 0.476), (u'paintings', 0.475), (u'painted', 0.4566), (u'canvas', 0.453), (u'canvases', 0.433), (u'artwork', 0.4223), (u'framed', 0.419), (u'ships', 0.411), (u'portrait', 0.396), (u'piece', 0.391)].

This indicates that if we did not have multiple-choice set available as in a factoid type questions the results would have been significantly worse.

We also compare coincidence of correct answers between corpora for the same question. This is illustrated in Table 4 for the first 9 questions (ones indicate the correct answer, zeros the incorrect one).

It is seen that the correct results for Glove Large and Glove Wiki coincide 2 times, and each of the corpora gives correct results 2 times when the other is incorrect. One could define a heuristic weighted distance.

$$d_{av} = ad_{GL} + (1 - \alpha)d_{GW} \tag{12}$$

Using $\alpha = 0.1$ gives the best result of 113 as a number of correct answers using word embedding only. Percentage wise we get 48.71 compared to 40.09 obtained in [26] with multi-sense.

5 Conclusions

This paper presents an improvement over the results of Wang et al. on Query Answering to IQ Test Questions using Word Embedding [26]. The improvement comes from using the Glove method and renormalization of results using the best results of two

leading methods. The latter approach combines knowledge coming from different corpuses. There are other possible ways to improve the results. First is the matrix reweighting. Second is adding documents from search engine that incorporate missing words both in questions and answers. If we used a training method we could apply formula (12) to each category, separately. Finally, careful studies should be done for finding regularities for each category of questions. It is not understandable why in [26] the multi-sense results on Antonyms are worse that without multi-sense.

We are convinced that one can achieve the result of 50 % of the correct answers without relational knowledge, compared to 40 % in [26].

Using the approach such as described in [17, 21] one could probably reach the result of accuracy approaching 60 %, compared with the best result of 50.09 % reported in [26]. This would be comparable with the best results achieved by humans. Still, a careful comparison with factoid type queries should be done. The further progress could be achieved but it would require a more standardized corpora than the ones that are available to date.

Acknowledgments This work was supported by the 04/45/DSPB/0136 PUT grant.

References

1. Besold, T., Hernąndez-Orallo, J., Schmid, U.: Can Machine Intelligence be Measured in the Same Way as Human intelligence? KI **29**(3): 291–297. Springer, Heidelberg (2015)
2. Bian, J., Chen, E., Dai, H., Gao, B., Liu, T.-Y., Tian, F., Zhang, R.: A probabilistic model for learning multi-prototype word embeddings. In: Proceedings of the 25th International Conference on Computational Linguistics (2014)
3. Bian, J., Gao, B., Liu, T.-Y.: Wordrep: a benchmark for research on learning word representations. CoRR (2014). arXiv:1407.1640
4. Blei, D.M., Jordan, M.I., Ng, A.Y.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Boguraev, B.K., Chu-Carroll, J., Fan, F., Fodor, P., Lally, A., McCord, M.C., Patwardhan, S., Prager, J.M.: Question analysis: How Watson reads a clue. IBM J. Res. Dev. **56**(3/4) (2012)
6. Bordes, A., Chopra, S., Mikolov, T., Weston, J.: Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. ICLR (2016)
7. Buckland, L.P., Voorhees, E.M.: The Sixteenth Text REtrieval Conference Proceedings. NIST Special Publication 500-274 (2007)
8. Bullinaria, J., Levy, J.: Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. Behav. Res. Methods **44**, 890907 (2012)
9. Carter, P.: The complete book of intelligence tests. John Wiley & Sons Ltd (2005)
10. Carter, P.: The Ultimate IQ Test Book: 1,000 Practice Test Questions to Boost Your Brain Power. Kogan Page Publishers (2007)
11. Chen, K., Corrado, G., Dean, J., Mikolov, T., Sutskever, I.: Distributed representations of words and phrases and their compositionality. CoRR (2013). arXiv:1310.4546
12. Dowe, D.L., Hernąndez-Orallo, J., Martnez-Plumed, F., Schmid, U., Siebers, M.: Computer models solving intelligence test problems: progress and implications. Artificial Intelligence **230**, 74–107 (2016)
13. Ferrucci, D.A.: Introduction to "This is Watson", IBM J. Res. Dev. **56**(3.4) ,11, IBM (2012)
14. Glove Large model. http://nlp.stanford.edu/data/glove.840B.300d.zip
15. Glove Small model. http://nlp.stanford.edu/data/glove.42B.300d.zip

16. Google News pretrained model. https://github.com/3Top/word2vec-api/blob/master/GoogleNews-vectors-negative300.bin.gz
17. Grishman, R., Nguyen, T.H.: Employing word representations and regularization for domain adaptation of relation extraction. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 6874. Baltimore (2014)
18. Huang, E.H., Manning, C.D., Ng, A.Y., Socher, R.: Improving word representations via global context and multiple word prototypes. In: Proceedings of ACL, 873-882 (2012)
19. http://ai.stanford.edu/~ehhuang/
20. Manning, C.D., Pennington, J., Socher R.: Glove:Global vectors for word representation. In: Proceedings of the Empirical Methods in Natural Language Processing (2014)
21. Moschitti, A., Plank, B.: Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 1: Long pp., 14981507. Sofia, Bulgaria (2013)
22. Prade, H., Richard, G.: From Analogical Proportion to Logical Proportions. Springer, Basel (2013)
23. Questions and answers on-line database. http://www.indiabix.com/
24. TAC 2008 Question Answering Track. http://www.nist.gov/tac/2008/qa/index.html
25. Verbal questions test set. http://research.microsoft.com/en-us/um/beijing/events/DL-WSDM-2015/VerbalQuestions.zip
26. Wang, H., Gao, B., Bian, J., Tian, F., Tie-Yan, L.: Solving verbal comprehension questions in IQ test by knowledge-powered word embedding. CoRR (2015). arXiv:1505.07909
27. Wechsler Adult Intelligence Scale. http://wechsleradultintelligencescale.com
28. Word2vec. https://code.google.com/p/word2vec

Identification of a Multi-criteria Assessment Model of Relation Between Editorial and Commercial Content in Web Systems

Jarosław Jankowski, Wojciech Sałabun and Jarosław Wątróbski

Abstract Together with the increasing role of Internet in commercial activity growing intensity of marketing content is observed. Advertising clutter is interfering with web usability and is affecting processing of the editorial content by web users. Therefore, effective way to manage marketing content is needed. This problem can be solved by using a proper combination of multi-criteria decision-analysis methods. The presented research shows a unique approach to identify assessment model of tradeoffs between the editorial content and the intensity of marketing components. The fuzzy model is identified on the basis of the experiment with the use of eye tracker and a combination of PROMETHEE and COMET methods. As a result, we obtained the assessment model, which is a relation between a set of defined inputs and a set of permissible outputs with the property that each input is related to exactly one output (assessment). Therefore, this model can be used online to manage web systems with balance between editorial and commercial content.

Keywords Web systems · MCDA · COMET · Fuzzy logic

J. Jankowski (✉) · W. Sałabun · J. Wątróbski
West Pomeranian University of Technology, Szczecin al. Piastów 17,
70-310 Szczecin, Poland
e-mail: jjankowski@wi.zut.edu.pl

W. Sałabun
e-mail: wsalabun@wi.zut.edu.pl

J. Wątróbski
e-mail: jwatrobski@wi.zut.edu.pl

J. Jankowski
Wrocław University of Technology, Wybrzeże Wyspiaskiego 27,
50-370 Wrocław, Poland

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_26

1 Introduction

Web presence is a key part of marketing strategies for more and more organisations. As a result design of Web systems requires taking into an account commercial goals as well as point of view of target users. Final applications should match target group's preferences and the efforts are taken to improve user experience and system usability with the use of evaluation methods or heuristics [4]. In platforms with substantial audience the exploitation of advertising space within editorial content delivers additional revenues and dedicated technologies are used to maximize results. Decision support systems for online campaign planning and real time optimization are used [3, 6]. Attempts to increase the performance are targeted to maximizing the profits and direct response but the business goals and intensive marketing are often conflicting with the user expectations. They often lead to negative side effects such as growing intrusiveness of online marketing content and increasing stimuli interrupts audience's cognitive processes [31]. Intrusiveness is considered as the degree to which it attempts to change users behaviour and negatively affects the user's experience [8]. As a result users are trying to avoid marketing content what results both physical and cognitive avoidance [7]. Searching for compromise between content intrusiveness and its influence on user experience within a web system is one of directions of research in this area [22]. Proposed in this paper approach used novel hybrid combination of two multi-criteria decision-analysis (MCDA) methods to identify assessment model.

The paper is an extension of previous work [11], where the PROMETHEE method was used to assess a set of decision-making alternatives. The considered problem presents an assessment of tradeoff between marketing content and the web user experience, where a big number of component criteria creates an area for the use the MCDA methods [28, 29, 32]. The main purpose of this paper is identifying a fuzzy model, which could be used repeatedly for different sets of variants in the whole domain of decision-making. In this way, if the set of alternatives is changed, then the model will further be used to make the assessment. The MCDA method has not to be repeated because instead we can use the identified model. It will be identified with use of the COMET method.

Thus, the main novelty of presented paper is identified the values of preferences, not only for selected set of alternatives, but for the all space of the considered problem. Therefore, the sensitivity of preference for each variant can be easily computed. It can be helpful for more effective management.

The rest of the paper is organized as follows. In Sect. 2, we give an outline of a literature review for design of online systems and their integration with marketing content. Sections 3 and describe the methodology used to identify the assessment model. Section 4 presents experimental study and results. Finally, we discuss results and their importance in Sect. 5.

2 Literature Review

Integration of commercial and editorial content takes place in most of online systems targeted to massive audiences. Design processes include several stages and decisions are related to the layout and integration of commercial goals with system usability. They should take into account factors related to all engaged parties i.e. advertisers and first of all the perspective of user experience. The effects of advertising content on web users are important not only for advertisers but for the strategy of web portal owners as well. From their point of view and the need for delivering websites with high usability important is tracking the quality and intensity of marketing content provided within their websites [21]. Earlier research reported several usability problems related to online advertising with misleading information and difficult to find options to remove advertising content [5]. Intensive online advertising can affect websites and more difficult information seeking, increased cognitive load may result frustration and other negative emotions [6]. Even though advertising content is delivered by external parties, negative altitudes can be developed towards website and as a result shorter time spent within website. Without reducing levels of intrusiveness in longer term, websites exploring extensively advertising space can suffer from dropping audiences and negative feedback [12]. At websites overloaded with advertising content users perceive the advertising clutter when non-editorial content exceeds acceptable level by users [10]. It brings negative results to advertisers with hard to remember marketing messages among competing content and is irritating web users making access to editorial content more difficult. As a results overall user experience within the website is negatively affected, what was analysed from the perspective of visual attention theories, oriented response and limited capacity model in relation to advertisements affecting information seeking performance within websites [20].

While online users are overloaded by different information, only part of content takes attention because of the limited ability to process information [14]. Attempts to acquire attention are connected with elements with high visibility with the usage of vivid effect and luminance [3]. Excessive usage of video, audio and animations causes overload problem of commercial content and is leading to side effects followed by affecting negatively user experience [23]. Overall effectiveness of campaign can be reduced because of limited cognitive capacity when negative affective response like irritation and annoyance takes place [30]. Overload with information on the web is resulting engaging in selective perception online and only limited number of messages is processed while other are ignored. Models predicting different orienting response and avoiding persuasive content as a function of content repetition within interactive environment to explain that phenomenon were proposed [22].

Advertisers try to evaluate the levels of intrusiveness and how it is affecting effectiveness of online campaigns. Research related to the forms of intrusive advertising deals with the different aspects of intrusiveness affecting information processing like frequency, sizes or ability to remove the content [9]. Attitude towards intrusive advertising was analysed from the perspective of mere exposure effect based on increased preference towards repeated stimuli. Authors showed that brands suffer

from intrusive advertising when is recognized by customers [16]. Too high intensity of animations and moving elements it resulting negative effect for aided recall during experiments. From advertisers point of view the influence of online content on brand recall and brand altitude can be crucial. Even though intensive advertising content is attracting attention, high intrusiveness lead to situation when different techniques to disable advertising content using physical avoidance and dedicated software are developed [13]. Cognitive and physical avoidance can be observed and theories from traditional media were extended towards better understanding that phenomena [2]. Apart from attentional avoidance side effects are observed and identified as banner blindness [1].

Usage of techniques attracting users attention may increase effectiveness of advertising in short term but overload and distraction of web users attention from the main editorial content result dropping number of users [15]. Proposed in this paper approach can be used for balanced approach with the ability to adjust the intensity of advertising and its influence on editorial content processing. It enables the acquisition of an acceptable level of effects represented by attention of web users without having excessive negative influence. In the next part, the general assumptions of the proposed approach and results of experiments are presented.

3 Methodological Background

The main goal of proposed approach is achieving compromise between editorial and commercial content within websites. Presented approach is based on the concept of the COMET decision support method free of the Rank Reversal phenomenon [18, 19]. In previous works, the accuracy of the COMET method was verified [17]. The formal notation of the COMET method should be shortly recalled [24–27].

In the first step the expert determines the dimensionality of the problem by selecting the number r of criteria, $C_1, C_2, ..., C_r$. Then, the set of fuzzy numbers for each criterion C_i is selected (1):

$$\begin{aligned} C_1 &= \{\tilde{C}_{11}, \tilde{C}_{12}, ..., \tilde{C}_{1c_1}\} \\ C_2 &= \{\tilde{C}_{21}, \tilde{C}_{22}, ..., \tilde{C}_{2c_1}\} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ C_r &= \{\tilde{C}_{r1}, \tilde{C}_{r2}, ..., \tilde{C}_{rc_r}\} \end{aligned} \tag{1}$$

where subscripts $c_1, c_2, ..., c_r$ are numbers of the fuzzy numbers respectively for criteria $C_1, C_2, ..., C_r$.

In the next step the characteristic objects (CO) are obtained by using the Cartesian Product of fuzzy numbers cores for each criteria as follows (2):

$$CO = C(C_1) \times C(C_2) \times ... \times C(C_r) \tag{2}$$

As the result, the ordered set of all CO is obtained (3):

$$\begin{array}{l} CO_1 = \{C(\tilde{C}_{11}), C(\tilde{C}_{21}), ..., C(\tilde{C}_{r1})\} \\ CO_2 = \{C(\tilde{C}_{11}), C(\tilde{C}_{21}), ..., C(\tilde{C}_{r2})\} \\ ..... \\ CO_t = \{C(\tilde{C}_{1c_1}), C(\tilde{C}_{2c_2}), ..., C(\tilde{C}_{rc_r})\} \end{array} \tag{3}$$

where t is the number of CO (4):

$$t = \prod_{i=1}^{r} c_i \tag{4}$$

Step three is based on ranking of the characteristic objects. The expert determines the Matrix of Expert Judgment (MEJ). It is a result of pairwise comparison of the COs by the problem expert. The MEJ matrix contains results of comparing characteristic objects by the expert,

$$MEJ = \begin{pmatrix} \alpha_{11} & \alpha_{12} & ... & \alpha_{1t} \\ \alpha_{21} & \alpha_{22} & ... & \alpha_{2t} \\ ... & ... & ... & ... \\ \alpha_{t1} & \alpha_{t2} & ... & \alpha_{tt} \end{pmatrix} \tag{5}$$

where α_{ij} is the result of comparing CO_i and CO_j by the expert. The function f_{exp} denotes the mental function of the expert. It depends solely on the knowledge of the expert and can be presented as (6). Afterwards, the vertical vector of the Summed Judgments (SJ) is obtained as follows (7).

$$\alpha_{ij} = \begin{cases} 0.0, f_{exp}(CO_i) < f_{exp}(CO_j) \\ 0.5, f_{exp}(CO_i) = f_{exp}(CO_j) \\ 1.0, f_{exp}(CO_i) > f_{exp}(CO_j) \end{cases} \tag{6}$$

where f_{exp} is the expert mental judgment function.

$$SJ_i = \sum_{j=1}^{t} \alpha_{ij} \tag{7}$$

Finally, values of preference are approximated for each characteristic object. As a result, the vertical vector P is obtained, where $i-th$ row contains the approximate value of preference for CO_i. The principle of an insufficient reason and SJ vector are used to this aim. The best CO gets an one point, and the worst gets a zero point.

In the next step each characteristic object and value of preference is converted to a fuzzy rule as follows (8). As a result, the complete fuzzy rule base is obtained.

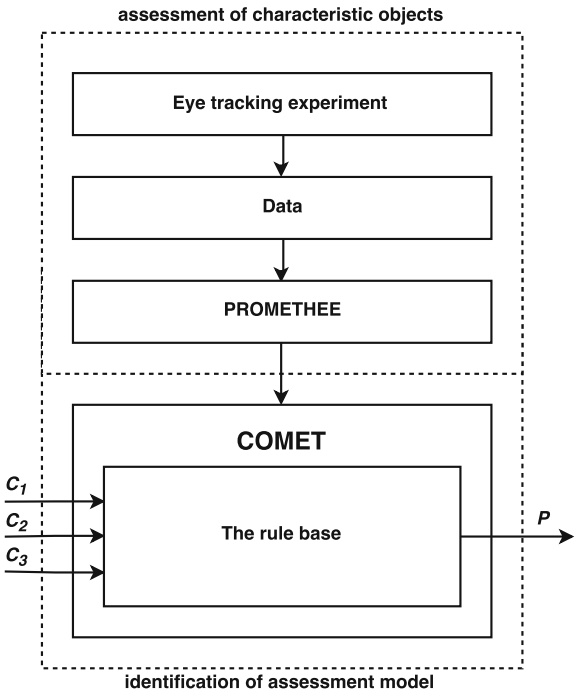


Fig. 1 Decision support system of web content assessment based on the eye tracking

$$IF\ CO_i\ THEN\ P_i \tag{8}$$

In the final step inference and final ranking each alternative is presented as a set of crisp numbers (e.g., $A_i = \{a_{1i}, a_{2i}, ..., a_{ri}\}$). This set corresponds to criteria C_1, C_2, ..., C_r. Mamdani's fuzzy inference method is used to compute preference of $i - th$ alternative. The rule base guarantees that the obtained results are unequivocal. The identified model is a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output. Therefore, the COMET a completely rank reversal free. Figure 1 presents the complete procedure of proposed approach.

4 Experimental Study and Results

The paper is an extension of previous work [11], where the PROMETHEE method was used to assess a set of decision-making alternatives. We used the previous results to propose a new approach to identifying the full domain model for assessment of tradeoff between marketing and editorial content with limited impact on the web user experience. The first step to identify the assessment model is based on defining the space of the problem. We select three criteria to model, i.e., intensity,

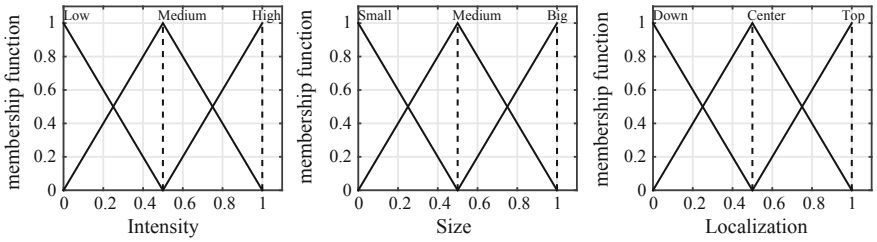


Fig. 2 Definitions of linguistic values for size (S), intensity (I) and localization (L)

Table 1 The set of characteristic objects in respect to criteria ($C_1 - C_3$), Summed Judgements (SJ) and Preferences (P)

CO_i	C_1	C_2	C_3	SJ	P
CO_1	Low	Small	Down	17.5	15/24
CO_2	Low	Small	Center	15.5	13/24
CO_3	Low	Small	Top	15.5	12/24
...	...	...	...	...	...
CO_{25}	High	Big	Down	0.5	0/24
CO_{26}	High	Big	Center	20.5	18/24
CO_{27}	High	Big	Top	11.5	9/24

size and localization of advertising content. Figure 2 presents three fuzzy levels of intensity $C_1 = \{low, medium, high\}$, size $C_2 = \{small, medium, big\}$ and localization $C_3 = \{down, center, top\}$. As a result, we get 27 characteristic objects as combinations of websites with integrated marketing content, which are presented in Table 1. The assessment of these objects can be difficult and problematic. Therefore, the PROMETHEE method and experiment were arranged to assess the characteristic objects described in detail in [11]. For each one characteristic object, the following set of factors from the experiment with the use eye tracker is obtained:

- the number of viewers to the marketing content (MV),
- percentage of total time spent on the website with the focus on advertisement (MTP),
- focus time on the marketing content (MT),
- time to the first view of the advertising content after the website is fully loaded (MFV),
- the number of repeated visits to the advertising content (MRV),
- the total number of returning visitors (MRVN),
- total time spent on editorial content (ET),
- the percentage representation of it (ETP),
- the number of viewers (EV),
- time to the first viewing of the editorial content (EFV),

Table 2 Three sample of alternatives and their assessment P

A_i	C_1	C_2	C_3	P
A_1	0.0677	0.7689	0.4579	0.6653
A_2	0.4372	0.1279	0.9279	0.2347
A_3	0.8929	0.5798	0.0373	0.7789

Table 3 Sensitivity for three alternatives assessment in respect to each criterion

A_i	$C_1 + 1\,\%$	$C_2 + 1\,\%$	$C_3 + 1\,\%$
A_1	−0.0837 %	+0.1165 %	+0.3867 %
A_2	−0.8709 %	+0.1055 %	−3.4641 %
A_3	−0.1976 %	−1.1838 %	−0.0453 %

- the number of revisits (ERV), and
- the number of returning visitors (ERVN).

On the basis of these data, the assessment of characteristic objects was made. First, the MEJ matrix is determined and then the vector SJ is calculated as sums of points for each characteristic object. In this way we get a ranking of characteristic objects, which is used to identification a assessment model for the entire domain, based on the principle of indifference of Laplace and the ranking of characteristic objects.

The summary of these steps is presented in Table 1 presents all characteristic objects in respect to three criteria, Summed Judgements (SJ) and values of Preference P. The final model is obtained by converted these data in respect to formulas (14), (15) and (16).

The identified model can be used to analysis sensitivity of a preference value. For this purpose, Table 2 presents a three simple alternatives and their level of preferences, which is calculated on the basis of obtained model. The assessment is very fast, because does not require anymore the participation of any expert. Therefore, it can be used to manage web systems with the focus on balance between editorial and commercial content.

Table 3 presents the simple analysis of sensitivity for three presented alternatives in respect to each criterion. It means how to change the rating if an input will be increase by one percent, e.g. if the value of third attribute for second alternative increase by 1 % then the assessment will be decrease by 3.4641 %. This approach requires compliance ceteris paribus assumption. On the basis of this simple experiment we can observed that selected criteria have not constant influence on preference values, e.g., increasing values of C_1 have caused decreasing of preference for all analyzed alternatives but it changes for criteria C_2 and C_3.

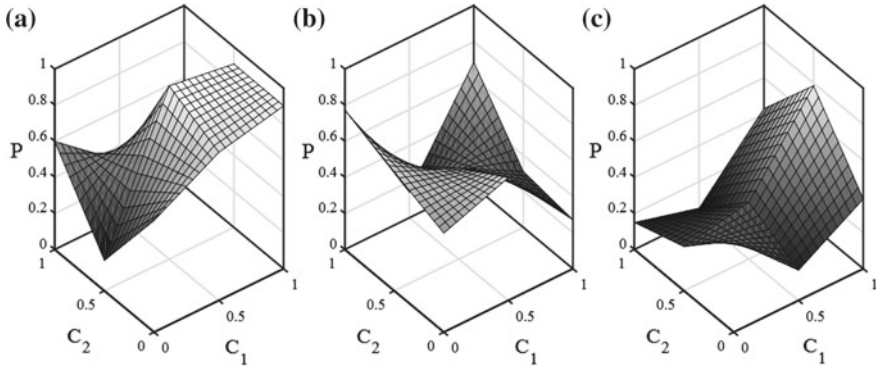


Fig. 3 Visualization of relationship between criteria C_1 and C_2 in respect to sample values of criterion C_3, **a** $C_3 = 0.0373$, **b** $C_3 = 0.4579$, **c** $C_3 = 0.9279$

Finally, the sensitivity of presented model can be visualized. For this purpose, we set one of inputs as the constant value. Then we can see the relationship between two remaining inputs. It is shown in Fig. 3. The all visualizations of relationships prove that weights of significance for each criterion is not constant and compensation is not linear.

5 Conclusions

The proposed approach is a useful tool from the MCDA domain, which can be applied in the complex problem as searching tradeoff between marketing content and the web user experience. The identified model can be used repeatable and it could be applied in the full space of the problem. Any number of alternatives or variants can be evaluated online because the values of the alternatives attributes do not affect the parameters of the model. This is the main reason why the COMET method is free of rank reversal phenomenon (change in the rank ordering when the set of alternatives has been changed).

Moreover, fuzzy modeling allows for use of varying levels of significance for each criterion. At the same time, it allows for use of relatively easy computation mechanism. Knowledge of the entire assessment model allows making sensitivity analysis of preferences for each criterion. Identified model can be the effective tool to manage website content in real time. Future studies should be focused on the possibility of determining local significance level of each criterion for enable a more complex analysis.

References

1. Benway, J.P., Lane, D.M.: Banner Blindness: Web Searchers Often Miss Obvious
2. Burke, R.R., Srull, T.K.: Competitive interference and consumer memory for advertising. J. Consum. Res. **15**, 55–68 (1988)
3. Du, H., Xu, Y.: Research on multi-objective optimization decision model of web advertising-takes recruitment advertisement as an example. Int. J. Adv. Comput. Tech. **4**(10), 329–336 (2012)
4. Flavian, C., Gurrea, R., Orus, C.: A heuristic evaluation of websites design for achieving the web success. Int. J. Serv. Stand. **5**(1), 17–41 (2008)
5. Gibbs, W.: Examining users on news provider web sites: a review of methodogy. J. Usab. Stud. **3**, 129–148 (2008)
6. Brajnik, G., Gabrielli, S.: A review of online advertising effects on the user experience. Int. J. Hum. Comput. Interact. **26**(10), 971–997 (2010)
7. Goldstein, D.G., McAfee, R.P., Suri, S.: The cost of annoying ads. In Proceedings of the 22nd International Conference on World Wide Web, pp. 459–470 (2013)
8. Ha, L.: Advertising clutter in consumer magazines: dimensions and effects. J. Adv. Res. **36**(4), 76–85 (1996)
9. Kalyanaraman, S., Ivory, J., Maschmeyer, L.: Interruptions and online information processing: the role of interruption type, interruption content, and interruption frequency. In: Proceedings of 2005 Annual Meeting of International Communication Association, pp. 1–32 (2005)
10. Ha, L., McCann, K.: An integrated model of advertising clutter in offline and online media. Int. J. Advers. **27**(4), 569–592 (2008)
11. Jankowski, J., Ziemba, P., Wątróbski, J., Kazienko, P.: Towards the tradeoff between online marketing resources exploitation and the user experience with the use of eye tracking. In Intelligent Information and Database Systems, pp. 330–343 (2016)
12. Jankowski, J., Wątróbski, J., Ziemba, P.: Modeling the impact of visual components on verbal communication in online advertising. Comput. Collect. Intel. pp. 44–53, (2015)
13. Krammer, V.: An effective defense against intrusive web advertising. In: Proceedings of the 2008 6th Annual Conference on Privacy, Security and Trust (PST'08), pp. 3–14 (2008)
14. Lang, A.: The limited capacity model of mediated message processing. J. Commun. **50**(1), 4670 (2000)
15. McCoy, S., Everard, A., Polak, P., Galletta, D.F.: The effects of online advertising. Commun. ACM **50**(3), 84–88 (2007)
16. Nielsen, J.H., Huber, J.: The Effect of Brand Awareness on Intrusive Advertising
17. Piegat, A., Sałabun, W.: Comparative analysis of MCDM methods for assessing the severity of chronic liver disease. In: International Conference on Artificial Intelligence and Soft Computing, pp. 228–238 (2015)
18. Piegat, A., Sałabun, W.: Identification of a multicriteria decision-making model using the characteristic objects method. Appl. Comput. Intel. Soft Comput. (2014)
19. Piegat, A., Sałabun, W.: Nonlinearity of human multi-criteria in decision-making. J. Theor. Appl. Comput. Sci. **6**(3), 36–49 (2012)
20. Ping, Z.: Pop-Up Animations: Impacts and Implications for Website Design and Online Advertising, HCI and MIS: Applications 5 (2006)
21. Pollifroni, M.: Multidimensional analysis applied to the quality of the websites: some empirical evidences from the italian public sector. Econ. Sociol **7**(4), 128–138 (2014)
22. Portnoy, F., Marchionini, G.: Modeling the effect of habituation on banner blindness as a function of repetition and search type: gap analysis for future work. In: Proceedings of the 28th of the International Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA'10), pp. 4297–4302 (2010)
23. Rosenkrans, G.: The creativeness and efectiveness of online interactive rich media advertising. J. Interact. Advers. **9**(2), (2009)
24. Sałabun, W.: Application of the fuzzy multi-criteria decision-making method to identify nonlinear decision models. Int. J. Comput. Appl. **89**(15), 1–6 (2014)

25. Sałabun, W.: Reduction in the number of comparisons required to create matrix of expert judgment in the comet method. Manag. Prod. Eng. Rev. **5**(3), 62–69 (2014)
26. Sałabun, W.: The characteristic objects method: a new distancebased approach to multicriteria decisionmaking problems. J. MCDA **22**(1–2), 37–50 (2015)
27. Sałabun, W.: The use of fuzzy logic to evaluate the nonlinearity of human multi-criteria used in decision making. P. Elektro. (Elec. Rev.) 88(10b), 235–238 (2012)
28. Wątróbski, J., Jankowski, J.: Knowledge Management in MCDA Domain. In: Proceedings of the FedCCSIS. Annual Computer Science and Information Systems 5, pp. 1445–1450 (2015)
29. Wątróbski, J., Jankowski, J.: Guideline for MCDA Method Selection in Production Management Area In: Rewski, P., Novikov, D., Bakhtadze, N., Zaikin, O. (eds.) New Frontiers in Information and Production Systems Modelling and Analysis. Intel. Syst. Refer. Lib. 98, pp. 119–138 (2016)
30. Yoo, CY., Kim, K.: Assessing the effects of animation in online banner advertising: hierarchy of effects model. J. Interact. Advers. **4**(2), 49-60
31. Zha, W., Wu, H.D.: The impact of online disruptive ads on users' com-prehension, evaluation of site credibility, and sentiment of intrusiveness. Am. Commun. J. **16**(2), 1528 (2014)
32. Ziemba, P., Piwowarski, M., Jankowski, J., Wątróbski, J.: Method of criteria selection and weights calculation in the process of Web projects evaluation. In: Hwang, D., Jung, J.J., Nguyen, N.-T. (eds.) ICCCI 2014. LNCS 8733, pp. 684–693 (2014)

Unsupervised Construction of Quasi-comparable Corpora and Probing for Parallel Textual Data

Krzysztof Wołk and Krzysztof Marasek

Abstract The multilingual nature of the world makes translation a crucial requirement today. Parallel dictionaries constructed by humans are a widely-available resource, but they are limited and do not provide enough coverage for good quality translation purposes, due to out-of-vocabulary words and neologisms. This motivates the use of statistical translation systems, which are unfortunately dependent on the quantity and quality of training data. Such systems have a very limited availability especially for some languages and very narrow text domains. Is this research we present our improvements to current quasi-comparable corpora mining methodologies by re-implementing the comparison algorithms, introducing a tuning script and improving performance using GPU acceleration. The experiments are conducted on lectures text domain and bi-data is extracted from web crawl from the WWW. The modifications made a positive impact on the quality and quantity of mined data and on the translation quality as well and used the BLEU, NIST and TER metrics. By defining proper translation parameters to morphologically rich languages we improve the translation quality and draw the conclusions.

Keywords SMT · Quasi comparable · Corpora · Parallel corpora generation · Comparable corpora · Unsupervised corpora acquisition · Data mining

K. Wołk (✉) · K. Marasek
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008 Warsaw, Poland
e-mail: kwolk@pja.edu.pl

K. Marasek
e-mail: kmarasek@pja.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_27

1 Introduction

The aim of this research is the preparation of parallel and quasi-comparable corpora and language models. This work improves SMT quality through the processing and filtering of parallel corpora and through extraction of additional data from the resulting quasi-comparable corpora. To enrich the language resources of SMT systems, adaptation and interpolation techniques will be applied to the prepared data. Experiments were conducted using data from a wide domain (TED[1] presentations on various topics).

Evaluation of SMT systems was performed on random samples of parallel data using automated algorithms to evaluate the quality and potential usability of the SMT systems' output [1].

As far as experiments are concerned, the Moses Statistical Machine Translation Toolkit software [2] is used. Moreover, the multi-threaded implementation of the GIZA++ tool is employed to train models on parallel data and to perform their symmetrization at the phrase level. The SMT system is tuned using the Minimum Error Rate Training (MERT) tool, which, through parallel data, specifies the optimum weights for the trained models, improving the resulting translations. The statistical language models from single-language data are trained and smoothed using the SRI Language Modeling toolkit (SRILM). In addition, data from outside the thematic domain is adapted. In the case of parallel models, Moore-Levis Filtering is used, while single-language models are linearly interpolated [3].

Lastly, methodology proposed in the Yalign [4] parallel data mining tool is analyzed and enhanced for the needs of quasi-comparable corpora exploration. Its speed is increased by re-implementing it in a multi-threaded manner and by employing graphics processing unit (GPU) horsepower for its calculations. Quality is improved by using the Needleman-Wunsch [5] algorithm for sequence comparison and by developing a tuning script that adjusts mining parameters to specific domain requirements.

The resulting systems out-performed baseline systems used in the tests.

2 Corpora Types

The term "parallel corpus" is typically used in linguistic circles to refer to texts that are translations of each other. A corpus is a large collection of texts, stored on a computer. Text collections are called corpora. For statistical machine translation, we are especially interested in parallel corpora, which are texts paired with a translation into another language. Texts differ in style and topic. For instance, transcripts of parliamentary speeches versus news reports. Preparing parallel texts for the purpose of statistical machine translation may require crawling the web,

[1] https://www.ted.com/.

extracting the text from formats such as HTML, and performing document and sentence alignment [3].

There are two main types of parallel corpora, which contain texts in two languages. In a comparable corpus, the texts are of the same kind and cover the same content. An example is a corpus of articles about football from English and Polish newspapers. In a translation corpus, the texts in one language (L1) are translations of texts in the second language (L2). It is important to remember that the term "comparable corpora" refers to texts in two languages that are similar in content, but are not translations of each other [3].

To exploit a parallel text, some kind of text alignment, which identifies equivalent text segments (approximately, sentences), is a prerequisite for analysis.

Machine translation algorithms for translating between a first language and a second language are often trained using parallel fragments, comprising a first language corpus and a second language corpus, which is an element-for-element translation of the first language corpus. Such training may involve large training sets that may be extracted from large bodies of similar sources, such as databases of news articles written in the first and second languages describing similar events. However, extracted fragments may be comparatively "noisy," with extra elements inserted in each corpus. Extraction techniques may be devised that can differentiate between "bilingual" elements represented in both corpora and "monolingual" elements represented in only one corpus, and for extracting cleaner parallel fragments of bilingual elements. Such techniques may involve conditional probability determinations on one corpus with respect to the other corpus, or joint probability determinations that concurrently evaluate both corpora for bilingual elements [3].

Because of such difficulties, high-quality parallel data is difficult to obtain, especially for less popular languages. Comparable corpora are the answer to the problem of lack of data for the translation systems for under-resourced languages and subject domains. It may be possible to use comparable corpora to directly obtain knowledge for translation purposes. Such data is also a valuable source of information for other cross-lingual, information-dependent tasks. Unfortunately, such data is quite rare, especially for the Polish–English language pair. On the other hand, monolingual data for those languages is accessible in far greater quantities [3].

Summing up, four main corpora types can be distinguished. Most rare parallel corpora can be defined as corpora that contain translations of the same document into two or more languages. Such data should be aligned, at least at the sentence level. A noisy parallel corpus contains bilingual sentences that are not perfectly aligned or have poor quality translations. Nevertheless, mostly bilingual translations of a specific document should be present in it. A comparable corpus is built from non-sentence-aligned and untranslated bilingual documents, but the documents should be topic-aligned. A quasi-comparable corpus includes very heterogeneous and non-parallel bilingual documents that may or may not be topic-aligned [6].

3 State of the Art

As far as comparable corpora are concerned, many attempts (especially for Wikipedia) have been made so far. Two main approaches for building comparable corpora can be distinguished. Perhaps the most common approach is based on the retrieval of cross-lingual information. In the second approach, source documents must be translated using any machine translation system. The documents translated in that process are then compared with documents written in the target language, to find the most similar document pairs.

An interesting idea for mining parallel data from Wikipedia was described in [7]. The authors propose two separate approaches. The first idea is to use an online machine translation (MT) system to translate Dutch Wikipedia pages into English, and then try to compare original EN pages with translated ones. The idea, although interesting, seems computationally infeasible, and it presents a chicken-egg problem. Their second approach uses a dictionary generated from Wikipedia titles and hyperlinks shared between documents. Unfortunately, the second method was reported to return numerous, noisy sentence pairs. The second method was improved in [8] by additional restrictions on the length of the correspondence between chunks of text and by introducing an additional similarity measure. They prove that [7] the precision (understood as number of correct translations pairs over total number of candidates) is about 21 %, and in the improved method [9], the precision is about 43 %.

Yasuda and Sumita [10] proposed an MT bootstrapping framework based on statistics that generate a sentence-aligned corpus. Sentence alignment is achieved using a bilingual lexicon that is automatically updated by the aligned sentences. Their solution uses a corpus that has already been aligned for initial training. They showed that 10 % of Japanese Wikipedia sentences have an English equivalent.

Another approach for exploring Wikipedia was recently described in [11] by Plamada and Volk. Their solution differs from the previously described methods in which the parallel data was restricted by the monotonicity constraint of the alignment algorithm used for matching candidate sentences. Their algorithm ignores the position of a candidate in the text and, instead, ranks candidates by means of customized metrics that combine different similarity criteria. In addition, the authors limit the mining process to a specific domain and analyze the semantic equivalency of extracted pairs. The mining precision in their work is 39 % for parallel sentences and 26 % for noisy-parallel sentences, with the remaining sentences misaligned. They also report an improvement of 0.5 points in the BLEU metric for out-of-domain data, and almost no improvement for in-domain data.

The authors in [12] propose obtaining only title and some meta-information, such as publication date and time for each document, instead of its full contents, to reduce the cost of building the comparable corpora. The cosine similarity of the titles' term frequency vectors was used to match titles and the contents of matched pairs.

In the research described in [13], the authors introduce a document similarity measure that is based on events. To count the values of this metric, they model documents as sets of events. These events are temporal and geographical expressions found in the documents. Target documents are ranked based on temporal and geographical hierarchies. The authors also suggest an automatic technique for building a comparable corpus from the web using news web pages, Wikipedia, and Twitter in [14]. They extract entities, time interval filtering, URLs of web pages, and document lengths as features for classification and for gathering the comparable data.

As presented above, most studies focus on extracting parallel sentences from noisy parallel corpora or comparable corpora, such as bilingual news articles, patent data or the Wikipedia. Few studies have been conducted on quasi–comparable corpora. Quasi–comparable corpora are available in far larger quantities than noisy parallel or comparable corpora, while the parallel sentence extraction task is significantly more difficult.

In [5] authors extend previous studies that treats parallel sentence identification as a binary classification problem. They proposes a novel classifier training method that simulates the real sentence extraction process. Furthermore, they uses linguistic knowledge of Chinese character features. Experimental results on quasi–comparable corpora indicate that this proposed approach performed good by reporting precision equal to 92 % and recall to 94 %.

Improvement to SMT of 1 BLEU point was reported in [9]. Authors proposed an accurate parallel fragment extraction system that used an alignment model to locate the parallel fragment candidates, and used an accurate lexicon filter to identify the truly parallel ones.

The [15] article addresses parallel data extraction from the quasi–parallel corpora generated in a crowd-sourcing project where ordinary people watch tv shows and movies and transcribe/translate what they hear, creating document pools in different languages. Since they do not have guidelines for naming and performing translations, it is often not clear which documents are the translations of the same show/movie and which sentences are the translations of the each other in a given document pair. Authors introduced a method for automatically pairing documents in two languages and extracting parallel sentences from the paired documents. The method consists of three steps document pairing, sentence pair alignment of the paired documents, and context extrapolation to boost the sentence pair coverage. Human evaluation of the extracted data shows that 95 % of the extracted sentences carry useful information for translation.

In [16] authors present a new implication of Wu's [14] Inversion Transduction Grammar (ITG) Hypothesis, on the problem of retrieving truly parallel sentence translations from large collections of highly non-parallel documents. The approach leverages a strong language universal constraint posited by the ITG Hypothesis, that can serve as a strong inductive bias for various language learning problems, resulting in both efficiency and accuracy gains. The described method introduced

exploits Bracketing ITGs to produce the first known results for this problem. Experiments showed that it obtained large accuracy gains of precision equal to 64.7 %.

In the present research, the method inspired by the Yalign tool is used. The solution was far from perfect, but after improvements and adaptation for the need of quasi-comparable corpora mining, that were made during this research, it supplied the SMT systems with bi-sentences of good quality in a reasonable amount of time.

4 Parallel Data Mining

In this research, methodologies that obtain parallel corpora from data sources that are not sentence-aligned and very non-parallel, such as quasi-comparable corpora, are presented. The results of initial experiments on text samples obtained from the Internet crawled data are presented. The quality of the approach used was measured by improvements in MT systems translations.

For the experiments in data mining, the TED corpora prepared for the IWSLT 2014 evaluation campaign by the FBK[2] was chosen. This domain is very wide and covers many unrelated subject areas. The data contains almost 2.5 M untokenized words [17]. The experiments were conducted on PL-EN (Polish-English) corpora.

The solution can be divided into three main steps. First, the quasi-comparable data is collected, then it is aligned based on keywords, and finally the aligned results are mined for parallel sentences. The last two steps are not trivial, because there are great disparities between polish-english documents. Text samples in English corpus are mostly misaligned, with translation lines whose placement does not correspond to any text lines in the source language. Moreover, most sentences have no corresponding translations in the corpus at all. The corpus might also contain poor or indirect translations, making alignment difficult. Thus, alignment is crucial for accuracy. Sentence alignment must also be computationally feasible to be of practical use in various applications.

Before a mining tool processes the data, it must be prepared. Firstly, all the data is downloaded and saved in a database. To obtain data a web crawler was developed. As input our tool requires bi-lingual dictionary or a phrase table with probability of such translation. Such input can be obtained using parallel corpora and tools like GIZA++ [18]. Based on such bi-lingual translation equivalents it is possible to make a query to the Google Search engine. Secondly we applied filtering of not likely translations and limiting number of crawled search results. In this research we used only 1-gram dictionary of words that for 70 % were translations of each other. For each keyword we crawled only 5 pages. Such strict limits were necessary because of time of web crawling, however it is obvious that much more data with better precision and domain adaptation will be obtained when

[2] http://www.fbk.eu/.

crawling higher order n-grams. In summary 43,899 pairs were used from the dictionary which produced almost 3.4 GB of data that contained 45,035,931 lines in english texts and 16,492,246 in polish texts. Average length of EN article was equal to 3,724,007 tokens and of PL to 4,855,009.

Secondly, our tool aligns article pairs and unifies the encoding of articles that do not exist in UTF-8. These keyword-aligned articles are filtered to remove any HTML tags, XML tags, or noisy data (tables, references, figures, etc.). Finally, bilingual documents are tagged with a unique ID as a quasi-comparable corpus. To extract the parallel sentence pairs, a decision was made to try strategy designed to automate the parallel text mining process by finding sentences that are close translation matches from quasi-comparable corpus. This presents opportunities for harvesting parallel corpora from sources, like translated documents and the web, that are not limited to a particular language pair. However, alignment models for two selected languages must first be created.

The solution was implemented using a sentence similarity metric that produces a rough estimate (a number between 0 and 1) of how likely it is for two sentences to be a translation of each other. It also uses a sequence aligner, which produces an alignment that maximizes the sum of the individual (per sentence pair) similarities between two documents [4].

For sequence alignment, the Yalign used an A* search approach [19] to find an optimal alignment between the sentences in two selected documents. The algorithm has a polynomial time worst-case complexity, and it produces an optimal alignment. Unfortunately, it cannot handle alignments that cross each other or alignments from two sentences into a single one [19].

After the alignment, only sentences that have a high probability of being translations are included in the final alignment. The result is filtered in order to deliver high quality alignments. To do this, a threshold value is used. If the sentence similarity metric is low enough, the pair is excluded.

For the sentence similarity metric, the algorithm uses a statistical classifier's likelihood output and normalizes it into the 0–1 range. The classifier must be trained in order to determine if sentence pairs are translations of each other. A Support Vector Machine (SVM) classifier was used in this research. Besides being an excellent classifier, an SVM can provide a distance to the separation hyperplane during classification, and this distance can be easily modified using a Sigmoid Function to return a likelihood between 0 and 1 [20].

The use of a classifier means that the quality of the alignment depends not only on the input but also on the quality of the trained classifier. To train the classifier, good quality parallel data were needed, as well as a dictionary that included translation probability. For this purpose, we used the TED talks [17] corpora. To obtain a dictionary, we trained a phrase table and extracted 1-grams from it [21].

5 Modifications to the Mining Process

Unfortunately, the native Yalign tool was not computationally feasible for large-scale parallel data mining and intended for comparable corpora exploration, that has higher parallelism level than quasi-comparable corpora. The standard implementation accepts plain text or web links, which need to be accepted, as input, and the classifier is loaded into memory for each pair alignment. In addition, the Yalign software is single-threaded. To improve performance, a solution was developed that supplies the classifier with articles from the database within one session, with no need to reload the classifier each time. The developed solution also facilitated multi-threading and decreased the mining time by a factor of 4.3 (using a 4-core, 8-thread i7 CPU). The alignment algorithm was also re-implemented for better accuracy and to leverage the power of GPUs for additional computing requirements. The tuning algorithm was implemented as well.

5.1 Needleman-Wunsch Algorithm (NW)

The objective of this algorithm is to align two sequences of elements (letters, words, phrases, etc.). The first step consists of defining the similarity between two elements. This is defined by the similarity matrix S, an N × M matrix, where N is the number of elements in the first sequence and M is the number of elements in the second sequence. The algorithm originated in the field of bioinformatics for RNA and DNA comparison. However, it can be adapted for text comparison. In simple terms, the algorithm associates a real number with each pair of elements in the matrix. The higher the number, the more similar the two elements are. For example, imagine that we have the similarity matrix S (phrase-polish, phrase-english) = number between 0 and 1. A 0 for two phrases means they have nothing in common; 1 means that those two phrases are the exact translation of each other. The similarity matrix definition is fundamental to the results of the algorithm [19].

The second step is the definition of the gap penalty. It is necessary in the case when one element of a sequence must be associated with a gap in the other sequence; however, such a step will incur a penalty (p).

The calculation of the M matrix is performed starting from the M0,0 element that is, by definition, equal to 0. After the first row and columns are initialized, the algorithm iterates through the other elements of the M matrix, starting from the upper-left side to the bottom-right side. Each step of this calculation is shown in Fig. 1.

The two NW algorithms, with and without GPU optimization, are conceptually identical, but the second has an advantage in efficiency, depending on the hardware, of up to max(n, m) times.

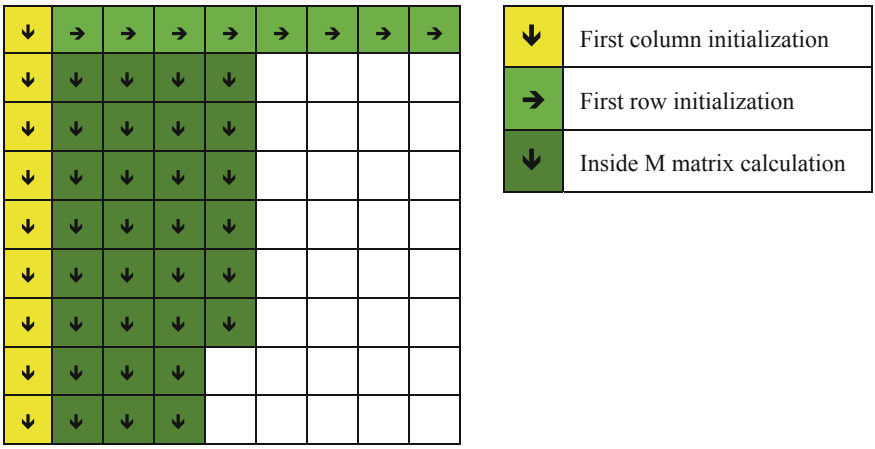


Fig. 1 Needleman-Wunsch M-matrix calculation

It differs in the calculation of the M matrix elements. This calculation is the step to which multi-threading optimization is applied. Those operations are small enough to be processed by an enormous number of Graphics processing units (ex. CUDA cores). The idea is to compute all elements in an anti-diagonal in parallel, always starting from the top-left and proceeding to the bottom-right. An example is presented in Fig. 2 [22].

Nonetheless, the results of the A* algorithm, if the similarity calculation and the gap penalty are defined as in the NW algorithm, will be the same only if there is an additional constraint on paths: paths cannot go upward or leftward in the M matrix. Yalign does not impose these additional conditions, so in some scenarios, repetitions of the same phrase may appear. In fact, every time the algorithm decides to move up or left, it is coming back into the second and first sequence respectively.

An example of an M matrix without constraints is presented in Table 1 and the same problem, the NW would react as presented in Table 2.

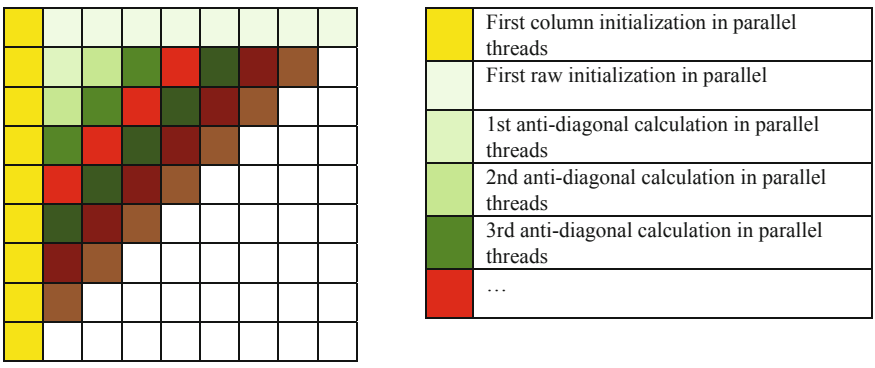


Fig. 2 Needleman-Wunsch M-matrix calculation with parallel threads

Table 1 Without constraints

	a	d	e	g	f
a	X				
d		X			
c	X				
d		X			
e			X	X	X

The alignment in this scenario is
a, d, a, d, e, g, f
a, d, c, d, e, –, –

Table 2 With NW

	a	d	e	g	f
a	X				
d		X			
c		X			
d		X			
e			X	X	X

The alignment using NW would be
a, d, –, –, e, g, f
a, d, c, d, e, –, –

Because of the lack of constraints, repetitions were created that visualized the imperfection of the A* algorithm implemented in the Yalign program. Using A* many sentences may be misaligned or missed during the alignment, especially when analyzed texts are of different lengths and have vocabularies rich in synonyms which exactly the case of quasi-comparable corpora. Some sentences can simply be skipped while checking for alignment. That is why NW with GPU optimization is the most suitable algorithm. In this research, a comparison was made using all three approaches described here.

5.2 Tuning Algorithm for Classifier

The quality of alignments is defined by a tradeoff between precision and recall. The classifier has two configurable variables:

- threshold: the confidence threshold to accept an alignment as "good." A lower value means more precision and less recall. The "confidence" is a probability estimated from a support vector machine classifying "is a translation" or "is not a translation."
- penalty: controls the amount of "skipping ahead" allowed in the alignment [4]. Say you are aligning subtitles, where there are few or no extra paragraphs and the alignment should be more or less one-to-one; then the penalty should be high. If you are aligning things that are moderately good translations of each

other, where there are some extra paragraphs for each language, then the penalty should be lower.

Both of these parameters are selected automatically during training but they can be adjusted if necessary. The solution implemented in this research also introduces a tuning algorithm for those parameters, which allows for better adjustment of them.

To perform tuning, it is necessary to extract random article samples from the corpus. Such articles must be manually aligned by humans. Based on such information, the tuning script tries, naively by random parameter selection, to find values for which classifier output is as similar to that of a human as possible. Similarity is a percentage value of how the automatically-aligned file resembles the human-aligned one. A Needleman-Wunsch algorithm is used for this comparison. For testing purposes, 25 random article pairs were taken from the quasi-comparable corpus and aligned by a human translator. Second, a tuning script was run using classifiers trained on the previously described text domain. A percentage change in quality was calculated for the classifier and equal to 7.3 %.

5.3 Evaluation of Improvements to the Tool

As mentioned, some methods for improving the performance of the native classifier were developed. First, speed improvements were made by introducing multi-threading to the algorithm, using a database instead of plain text files or Internet links, and using GPU acceleration in sequence comparison. More importantly, two improvements were made to the quality and quantity of the mined data. The A* search algorithm was modified to use Needleman-Wunsch, and a tuning script of mining parameters was developed. In this section, the PL-EN TED corpus will be used to demonstrate the impact of the improvements (it was the only classifier used in the mining phase). The data mining approaches used were: directional (PL- > EN classifier) mining (MONO), bi-directional (additional EN- > PL classifier) mining (BI), bi-directional mining using a GPU-accelerated version of the Needleman-Wunsch algorithm (NW), and mining using the NW version of the classifier that was tuned (NWT). The results of such mining are shown in Table 3.

As presented in Table 4, each of the improvements increased the number of parallel sentences discovered. In addition, in Table 4 a speed comparison is made using different versions of the tool. A total of 1,000 comparable articles were randomly selected from Wikipedia and aligned using the native implementation

Table 3 Obtained bi-sentences

Mining method	Number of bi-sentences
MONO	21,132
BI	23,480
NW	24,731
NWT	27,723

Table 4 Computation time

Mining method	Computation time (s)
Y	272.27
MY	63.1
NWMY	112.6
GNWMY	74.4

(Y), multi-threaded implementation (MY), classifier with the Needleman-Wunsch algorithm (NWMY), and with a GPU-accelerated Needleman-Wunsch algorithm (GNWMY).

The results indicate that multi-threading significantly improved speed, which is very important for large-scale mining. As anticipated, the Needleman-Wunsch algorithm decreases speed. However, GPU acceleration makes it possible to obtain performance almost as fast as that of the multi-threaded A* version. It must be noted that the mining time may significantly differ when the alignment matrix is big (text is long). The experiments were conducted on a hyper-threaded Intel Core i7 CPU and a GeForce GTX 980 GPU.

6 Evaluation of Obtained Comparable Corpora

Using techniques described above, we were able to build quasi-comparable corpora and mine it for parallel sentences for the Polish-English language pair and evaluate it using data being part of IWSLT 2014 conference. The obtained parallel corpus (EXT) and the TED corpora statistics are presented in Table 5.

To evaluate the corpora, we trained baseline systems using IWSLT 2014 official data sets and enriched them with obtained quasi-comparable corpora, both as parallel data and as language models. The enriched systems were trained with the baseline settings but additional data was adapted using linear interpolation and Modified Moore-Levis [23]. Because of the well know MERT instability, tuning was not performed in the experiments [24].

The evaluation was conducted using official test sets from IWSLT 2010–2013 campaigns and averaged. For scoring purposes, Bilingual Evaluation Understudy (BLEU) metric was used. The results of the experiments are shown in Table 6. BASE in Table 6 stands for baseline system and EXT for enriched systems.

As anticipated, additional data sets improved overall translation quality for each language and in both translation directions. The gain in quality was observed mostly in the English to foreign language direction.

Table 5 Corpora statistics

Corpus	Number of bi-sentences	Number of unique EN tokens	Number of unique foreign tokens
TED	2,459,662	2,576,938	2,864,554
EXT	818,300	1,290,000	1,120,166

Table 6 Results of MT experiments

Lang	System	Direction	Bleu
PL-EN	BASE	→ EN	30.21
	EXT	→ EN	31.37
	BASE	EN←	21.07
	EXT	EN←	22.47

7 Conclusions

Bi-sentence extraction has become more and more popular in unsupervised learning for numerous specific tasks. This method overcomes disparities between English and other languages. It is a language-independent method that can easily be adjusted to a new environment, and it only requires parallel corpora for initial training. Our experiments show that the method performs well. The resulting corpora increased MT quality in a wide text domain. Even small differences can make a positive influence on real-life, rare translation scenarios. In addition, it was proven that mining data using two classifiers trained from a foreign to a native language and vice versa, can significantly improve data quantity, even though some repetitions are possible. From a practical point of view, the method requires neither expensive training nor language-specific grammatical resources, but it produces satisfying results. It is possible to replicate such mining for any language pair or text domain, or for any reasonable comparable input data.

Acknowledgments This research was supported by Polish-Japanese Academy of Information Technology statutory resources (ST/MUL/2016), resources for young researchers at PJATK and CLARIN ERIC research program.

References

1. Wołk, K., Marasek, K.: Real-time statistical speech translation. New Perspectives in Information Systems and Technologies, vol. 1, pp. 107–113. Springer International Publishing (2014)
2. Wołk, K., Marasek, K.: Polish–English speech statistical machine translation systems for the IWSLT 2013. In: Proceedings of the 10th International Workshop on Spoken Language Translation, Heidelberg, Germany, pp. 113–119 (2013)
3. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2009)
4. Berrotarán, G., Carrascosa, R., Vine, A.: Yalign documentation. Accessed 01 2015

5. Chu, C., Nakazawa, T., Kurohashi, S.: Chinese-Japanese parallel sentence extraction from quasi–comparable corpora. ACL **2013**, 34 (2013)
6. Wu, D., Fung, P.: Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. Natural Language Processing—IJCNLP 2005. Lecture Notes in Computer Science, vol. 3651, pp. 257–268 (2005)
7. Adafree, S.F., deRijke, M.: Finding similar sentences across multiple languages in Wikipedia (2006)
8. Mohammadi, M., and Aghaee, N.Q.: Building bilingual parallel corpora based on Wikipedia (2010)
9. Chu, C., Nakazawa, T., Kurohashi, S.: Accurate parallel fragment extraction from quasi–comparable corpora using alignment model and translation lexicon. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 1144–1150 (2013)
10. Yasuda, K., Sumita, E.: Method for building sentence-aligned corpus from Wikipedia (2008)
11. Plamada, M., Volk, M.: Mining for domain-specific parallel texts from the Wikipedia (2013)
12. Aker, A., Kanoulas, E., Gaizauskas, R.J., A light way to collect comparable corpora from the Web. LREC (2012)
13. Strötgen, J., Gertz, M., Junghans, C.: An event-centric model for multilingual document similarity. In: SIGIR'11: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, pp. 953–962 (2011)
14. Wu, D.: Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Comput. Linguist. **23**(3), 377–403 (1997)
15. Sarikaya, R., Maskey, S., Zhang, R., Jan, E. E., Wang, D., Ramabhadran, B., Roukos, S.: Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In: INTERSPEECH, pp. 432–435 (2009)
16. Wu, D., Fung, P.: Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In: Natural Language Processing–IJCNLP 2005, pp. 257–268 (2005)
17. Cettolo, M., Girardi, C., Federico, M.: WIT3: Web inventory of transcribed and translated talks. In: Proceedings of EAMT, Trento, Italy, pp. 261–268 (2012)
18. Bojar, O., Rosa, R., Tamchyna, A.: Chimera–three heads for English-to-Czech translation. In: Proceedings of the Eighth Workshop on Statistical Machine Translation. Association for Computational Linguistics Sofia, Bulgaria, pp. 90–96 (2013)
19. Musso, G.: Sequence alignment (Needleman-Wunsch, Smith-Waterman). http://www.cs.utoronto.ca/~brudno/bcb410/lec2notes.pdf
20. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. Lect. Notes Comput. Sci. **1398**(1998), 137–142 (2005)
21. Wołk, K., Marasek, K.: A sentence meaning based alignment method for parallel text corpora preparation. Advances in Intelligent Systems and Computing, vol. 275, pp. 107–114. Springer, Madeira Island, Portugal (2014). ISSN 2194-5357. ISBN 978-3-319-05950-1
22. Roessler R.: A GPU implementation of Needleman-Wunsch. Specifically for use in the Program PyroNoise 2 (2010)
23. Koehn, P., Haddow, B.: Towards effective use of training data in statistical machine translation. In: WMT'12 Proceedings of the Seventh Workshop on Statistical Machine Translation, Stroudsburg, PA, USA, 317–321 (2012)
24. Clark, J.H., Dyer, C., Lavie, A., Smith, N.A.: Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2, pp. 176–181. Association for Computational Linguistics (2011)

Active Learning-Based Approach for Named Entity Recognition on Short Text Streams

Cuong Van Tran, Tuong Tri Nguyen, Dinh Tuyen Hoang, Dosam Hwang and Ngoc Thanh Nguyen

Abstract The named entity recognition (NER) problem has an important role in many natural language processing (NLP) applications and is one of the fundamental tasks for building NLP systems. Supervised learning methods can achieve high performance but they require a large amount of training data that is time-consuming and expensive to obtain. Active learning (AL) is well-suited to many problems in NLP, where unlabeled data may be abundant but labeled data is limited. The AL method aims to minimize annotation costs while maximizing the desired performance from the model. This study proposes a method to classify named entities from Tweet streams on Twitter by using an AL method with different query strategies. The samples were queried for labeling by human annotators based on query by committee and diversity-based querying. The experiments evaluated the proposed method on Tweet data and achieved promising results that proved better than the baseline.

Keywords Named entity recognition · Active learning · Query strategy · Text streams

C. Van Tran · T.T. Nguyen · D.T. Hoang · D. Hwang (✉)
Department of Computer Engineering, Yeungnam University, Gyeongsan, South Korea
e-mail: dosamhwang@gmail.com

C. Van Tran
e-mail: vancuongqbuni@gmail.com

T.T. Nguyen
e-mail: tuongtringuyen@gmail.com

D.T. Hoang
e-mail: hoangdinhtuyen@gmail.com

N.T. Nguyen
Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wrocław, Poland
e-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_28

1 Introduction

Named entity recognition also known as entity identification and entity extraction is a subtask of information extraction. It identifies entities in documents and classifies them into predefined categories such as person names, locations, organizations, etc. [1, 16]. The NER problem plays an important role in many NLP applications and is a fundamental task of NLP systems. The extracted named entities can be utilized for various purposes such as entity relation extraction, document summarization [10, 15], speech recognition [9], and term indexing in information retrieval systems [3].

Many different machine learning approaches such as maximum entropy, hidden Markov models, support vector machines and conditional random fields (CRF) have been adopted for NER and achieved high accuracy based on a large annotated corpus. The supervised learning methods achieve high performance if they are applied to well-formatted text. However, achievement results are not as expected when applied to short and noisy messages. Twitter is a social network service that provides access to large volumes of data in real-time, but it is notoriously noisy and hard to tackle in NLP problems. For example, performance by the Stanford NER that uses the CRF model to train a classifier for CoNLL03 data dropped from 90.8 % to 45.8 % when it was applied to Tweets [8]. The length of a Tweet is 140 characters at most and Tweets contain different kinds of information, such as text, hyperlinks, user mentions, and hashtags. In addition, users often write Tweets with a freestyle and acronyms, and do not include extra information to explain the author's opinion. Another challenge for NLP systems is the large volume and the dynamic content in terms of time [5, 14]. The data from Twitter could be fed into processing systems as a data stream.

The unlabeled data are often easily obtained; however, annotating these texts can be rather tedious and time-consuming. The shortage of labeled data is an obstacle to supervised learning methods in developing application systems. Active learning is an attractive technique that addresses the shortage of labeled data for the training phase. Instead of training that relies on randomly labeled samples from a large corpus, the AL method chooses samples to label via optimal algorithms. Using different strategies, AL may determine a much smaller and the most informative subset from a large amount of unlabeled data. The motivation of this work is to query the most informative samples from Tweet streams using the AL method. This paper presents an AL method for classifying named entities from Tweet streams with different query strategies: query by committee and diversity-based querying. First, two classifiers trained on the CRF model and the maximum entropy model are used to classify unlabeled data in an arrival stream, and then they select dissimilar results in order to ask a human annotator to correct them. Second, the Tweets contain proper nouns in which the context is the least similar when compared to the existing training data that are queried for labeling. As a case study, experiments were conducted on Tweet data to assess proposed strategies. The method greatly reduced the training data and achieved results better than the random sampling.

The organization of this paper is as follows. Section 2 briefly presents related works. An introduction to the AL method is given in Sect. 3. Section 4 presents the proposed system, and the results are analyzed in the subsequent section. The conclusion and future work are presented in Sect. 6.

2 Related Work

Named entity recognition has attracted more interest from researchers in recent years, especially the problem recognizes named entities in microtexts, such as Tweets on Twitter. The first work to mention here was contributed by Ritter et al. [12]. They rebuilt an NLP tool beginning with parts of speech tagging. The NER system leverages the redundancy inherent in Tweets to achieve high performance by using labeled latent Dirichlet allocation to exploit freebase dictionaries in a semi-supervised learning method. Another approach was described by Liu et al. [8], who proposed combining the K-nearest neighbors algorithm with a linear CRF model in a semi-supervised learning method. Li et al. [7] also proposed a novel two-step unsupervised NER approach to recognizing named entities in Twitter data based on gregarious properties of named entities in a targeted Tweet streams. The method deals with streams, however, it does not determine the class of the identified entity, determining only if a phrase is an entity or not.

Yao et al. [17] presented an alternative AL strategy and combined this method with semi-supervised learning to reduce the labeling effort for a Chinese NER task. They utilized a strategy based on information density for the sample selection in a sequential labeling problem, which is suitable for both AL and self-training. They achieved an F1 score of 77.4 % with the proposed hybrid method on a Sighan bake-off 2006 Microsoft research of Asia NER corpus. Chen et al. developed and evaluated AL methods for a clinical NER task to identify concepts of medical problems and treatments from clinical notes [4]. They simulated AL experiments using a number of existing and novel algorithms in three different categories. Based on the learning curves of F1 score and the number of sentences, uncertainty sampling algorithms outperformed all other methods, and most diversity-based methods also performed better than random sampling. In another study [6], Hassanzadeh and Keyvanpour presented a variance-based AL method for the NER task that chooses informative entities to minimize the variance of the classifier currently built from labeled data. By finding entities where labeling by the current model was certain, they used self-training to resolve unlabeled samples. The experiments, when applied to the CoNLL03 English corpus showed that the method used considerably fewer numbers of manually labeled samples to produce the same results as when samples were selected in a random manner.

This paper presents an AL method to extract named entities from Tweet streams on Twitter. The proposed method is an effective way to solve the NER task on Twitter.

3 Active Learning-Based Approach

3.1 Overview

Active learning is a supervised learning method in which the learner controls the selection of necessary data for the learning phase. The key issue is how to recognize necessary data in order to ask the human annotator to annotate them. The learner will ask an expert in the related domain about labels of samples for which the learned model has made unreliable predictions so far. The main purpose of AL is to create as good a classifier as possible without supplying more labeled samples and more human effort in annotating data. Active learning has been successfully applied to a number of NLP tasks such as information extraction, named entity recognition, text categorization, and so on [11].

3.2 Scenarios

Different scenarios that have considered in the literature for the learner to make queries are (i) membership query synthesis, (ii) stream-based selective sampling, and (iii) pool-based sampling [13].

(i) Membership query synthesis: The learning system asks whether a particular domain sample belongs to the unknown concept or not. The learning system may request the labels for any unlabeled sample in the input space, including queries that the learning system generates de novo, rather than those sampled from some underlying natural distribution.

(ii) Stream-based selective sampling: The stream-based AL is important when data is continuously available and cannot be easily stored. The data can first be sampled from the actual distribution one at a time from the data source, and then the learning system can decide whether to request its label or not. If the input distribution is uniform, selective sampling may well behave like membership query learning. However, if the distribution is non-uniform and unknown, it is guaranteed that queries will still be sensible since they come from a real underlying distribution.

(iii) Pool-based sampling: Assume there is a huge pool of unlabeled data, which is usually assumed to be closed and queries select samples from this pool. The samples are queried according to an informativeness-measure technique, evaluating all samples in the pool. There is a main difference between this scenario and stream-based selective sampling. Stream-based selective sampling queries samples in a sequential way at a time the data arrives, and makes the query decisions individually, whereas pool-based sampling queries on the pool of available samples; therefore, it can evaluate the entire data set before selecting the most informative samples.

Algorithm 1 Active learning algorithm

Input: T - Time interval
Output: C - Classifiers

1: Label initial samples - L
2: Train initial classifiers C on L
3: **repeat**
4: Get arrival data - U
5: Preprocessing U
6: Apply query strategies to unlabeled data U to obtain D
7: Ask the annotator for labeling samples in D to obtain D'
8: Add D' to L
9: Retrain classifiers C on L
10: **until** Out of T or Annotator stops
11: **return** C

4 System Descriptions

4.1 System Procedure

This section presents a brief description of the proposed method. This work experiments with a practical stream-based AL scenario. The workflow of the idea is illustrated in Fig. 1, and the algorithm for the training phase is described in Algorithm 1. Some main steps in the algorithm are as follows.

Initial phase. The model parameters and data are initialized for the system. Initially, a set of random samples is selected to annotate as initial training data for the classification models. Two models are utilized to train classifiers: the CRF model and the maximum entropy model, respectively. This work uses the CRF model provided by Stanford[1] and the maximum entropy model provided by OpenNLP.[2]

Preprocessing. Elements such as mentions, hyperlinks, hashtags, and symbols are removed from Tweets.

Querying. To increase the training dataset, new samples are selected from unlabeled arrival data based on two query strategies: query by committee and diversity-based querying. The selected samples satisfy some criterion such as the classification results of models are dissimilar or the context of the proper noun is the least similar to the training data. The queried samples are labeled by the human annotator and then added to the training data to retrain models.

Training. The classification models are applied to the updated training data to retrain the classifiers.

[1] http://nlp.stanford.edu/software/CRF-NER.shtml.

[2] https://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html.

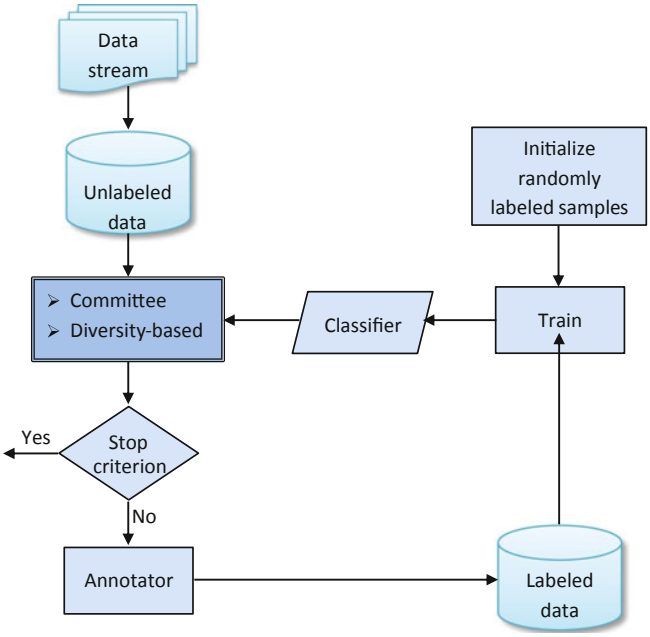


Fig. 1 The workflow of the system for the training phase

Iteration. The sampling and labeling processes are iterated until the stopping criterion is met. In this experiment, the system's processes are finished when the annotator stops working or they are beyond the considered time interval.

4.2 Context Features

A part of speech (POS) tagger assigns the parts of speech to each word or token, such as noun, verb, adjective, etc. The proper nouns are extracted based on the results of POS tagger (i.e., the experiments use a publicly available POS tagger developed by Stanford[3]).

With each proper noun in the unlabeled data and the named entity in the labeled data, a contextual vector is constructed to represent its contextual information. The size of the contextual vector is the size of the considered window. In the experiments, the window size is set to six (i.e., three words on the left and three words on the right of the proper noun or the named entity are examined). The elements of the vector are the POS tag of words in the considered window. Proper noun *PN* is represented by the contextual vector of the POS tag as follows:

[3] http://nlp.stanford.edu/software/.

$$PN = (..., pos_{-3}, pos_{-2}, pos_{-1}, pos_{+1}, pos_{+2}, pos_{+3}, ...) \quad (1)$$

where pos_i is the POS tag of the word at location i from the proper noun. The negative sign and the positive sign, respectively, mean that the words are to the left-hand side and to the right-hand side of the proper noun.

4.3 Query Strategies

Query by committee is a multi-classifier approach, where the classifiers are trained on the current training data by individual models (i.e., the CRF model and the maximum entropy model) and then they are used to examine the unlabeled arrival data. The disagreement between classifiers with respect to the value and the category of a named entity is utilized to decide whether the sample is to be labeled by the human annotator.

Diversity-based querying selects samples where the training data are the least similar. A vector model is used to measure the similarity between a Tweet and all the training data. The Tweets that contain a proper noun are examined for the similar context between the proper noun and named entities that exist in the current training data. Each proper noun and named entity is represented by a contextual vector, as presented above. The contextual vectors of proper nouns are then compared with all contextual vectors of the named entities in the training data. The Tweets where similar scores are less than a certain threshold will be subjected to labeling by the human annotator.

5 Experimental Results

5.1 Dataset and Baselines

The performance of the proposal method was evaluated by applying the system to Tweet data. The data consisted of Tweets of 20 users that were collected on Twitter from January 1st, 2014 to September 30th, 2015, by using the public Java library for the Twitter API[4] and then dropping Tweets with only hashtags, mentions, hyperlinks, or emoticons. Finally, 10,813 unlabeled Tweets were selected. In addition, 4,716 labeled Tweets were used as initial labeled data. The test set (*TS*) included 1,153 Tweets that were also collected on Twitter from October 1st, 2015 to December 31th, 2015, and annotated as the gold standard (*GS*) to assess the performance. The dataset was annotated with three named entity categories: Person, Location, and Organization.

[4] http://twitter4j.org.

The two systems implemented are the query by committee (QBC) algorithm and the diversity-based querying (DBQ) algorithm. A random sampling (RS) algorithm that presents a passive learning method was also implemented as a baseline compared to our proposed method.

5.2 Evaluation and Results

The performance of this task was calculated following #MSM2013's measures [2]. Precision, Recall, and F-measure are calculated for each entity category, and the final results for all entity categories are the average performance of the defined categories.

The entity is represented in a tuple (entity value, entity category), and the strict matching is performed between named entities in the *TS* and answers in the *GS* for both detection of the correct entity value and the correct entity category.

Precision ($\overline{P}$) and Recall ($\overline{R}$) for all entity categories are the average value of the precision and the recall of all entity categories, respectively. The F-measure score (also called F_1 score) is the harmonic mean of $\overline{P}$ and $\overline{R}$, defined as follows:

$$F_1 = 2 \times \frac{\overline{P} \times \overline{R}}{\overline{P} + \overline{R}} \tag{2}$$

The results for each entity category and overall results from systems are shown in Table 1. All systems were tested in the same AL framework (i.e., the same initial training data, model, default parameters for models, time interval, and test set). The experiments used the CRF model for testing data and set the time interval at 20 (i.e., the Tweets were queried during 20 days for each iteration).

The performance of all query methods (i.e., QBC and DBQ) outperformed the RS method, in which the selected Tweets of the AL method were less than or equal to the RS method. The F_1 score of the QBC was the best, where the F_1 score was 64 %.

Table 1 The performance of the systems

	System	Person	Location	Organization	All
Precision	*QBC*	**83.8**	83.3	78.6	81.9
	DBQ	79.7	**85.6**	**80.8**	**82**
	RS	79.5	83.3	66.7	76.5
Recall	*QBC*	59.6	**68.9**	**28.9**	**52.5**
	DBQ	**60.3**	67.7	27.6	51.9
	RS	57.1	65.9	23.7	48.9
F_1	*QBC*	**69.7**	75.4	**42.3**	**64**
	DBQ	68.6	**75.6**	41.2	63.6
	RS	66.4	73.6	35	59.7

Table 2 The number of selected tweets of each system

System	#Selected Tweets
QBC	2,165 (20.02 %)
DBQ	3,252 (30.1 %)
RS	3,252 (30.1 %)

That was better than the baseline 4.3 %. The QBC outperformed the DBQ since it required fewer labeled Tweets than the DBQ. The F_1 score of the two query methods are very similar (i.e., 64 % for QBC and 63.6 % for DBQ).

5.3 Discussions

The number of selected Tweets at the end of the training process from the AL method for each query method is shown in Table 2. Although the QBC method only selects 20.02 % of the unlabeled Tweets, its performance is better than the DBQ and the RS, which select 30.1 % of unlabeled Tweets.

Comparing the performance and the number of selected Tweets among the query strategies of the AL method, and between the AL method and the passive learning method, the QBC method is the best (i.e., to achieve a 64 % F_1 score, QBC queried 2,165 Tweets; DBQ queried 3,252 Tweets to achieve 63.6 %, and RS also queried 3,252 Tweets to achieve 59.7 %).

One concern with applying the QBC method to real tasks is that it relies on updated models, which require more time for training. In the experiments, it takes several minutes to fully train a model; this is also suitable in reality for the stream-based scenario. The DBQ method does not depend on trained models, but in the experiments, it did not outperform the QBC method.

6 Conclusion and Future Work

The active learning method aims to minimize annotation costs while maximizing the performance of the model. This study conducted AL experiments for NER with Tweet streams from Twitter and showed that the AL method has the potential to reduce annotation costs to train the model. By using two different query strategies (query by committee and diversity-based querying), the AL method achieved the performance better than the passive learning method.

Future work includes improving the current query strategies and proposing more query methods for the NER problem on Twitter.

Acknowledgments This work was supported by the BK21+ program of the National Research Foundation (NRF) of Korea.

References

1. Abdallah, S., Shaalan, K., Shoaib, M.: Integrating rule-based system with classification for arabic named entity recognition. In: Computational Linguistics and Intelligent Text Processing, pp. 311–322. Springer (2012)
2. Cano Basave, A.E., Varga, A., Rowe, M., Stankovic, M., Dadzie, A.S.: Making sense of microposts (#msm2013) concept extraction challenge (2013)
3. Chen, H.H., Ding, Y.W., Tsai, S.C.: Named entity extraction for information retrieval. Comput. Process. Orient. Lang. **12**(1), 75–85 (1998)
4. Chen, Y., Lasko, T.A., Mei, Q., Denny, J.C., Xu, H.: A study of active learning methods for named entity recognition in clinical text. J. Biomed. Inf. **58**, 11–18 (2015)
5. Giao, B.C., Anh, D.T.: Similarity search for numerous patterns over multiple time series streams under dynamic time warping which supports data normalization. Vietnam J. Comput. Sci. pp. 1–16 (2016)
6. Hassanzadeh, H., Keyvanpour, M.: A variance based active learning approach for named entity recognition. In: Intelligent Computing and Information Science, pp. 347–352. Springer (2011)
7. Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., Lee, B.S.: Twiner: named entity recognition in targeted twitter stream. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 721–730. ACM (2012)
8. Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1. pp. 359–367. Association for Computational Linguistics (2011)
9. Meyer, C., Schramm, H.: Boosting hmm acoustic models in large vocabulary speech recognition. Speech Commun. **48**(5), 532–548 (2006)
10. Nobata, C., Sekine, S., Isahara, H., Grishman, R.: Summarization system integrated with named entity tagging and ie pattern discovery. In: Proceedings of Third International Conference on Language Resources and Evaluation, pp. 1742–1745 (2002)
11. Olsson, F.: A literature survey of active machine learning in the context of natural language processing (2009)
12. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534. Association for Computational Linguistics (2011)
13. Settles, B.: Active learning literature survey. Univ. Wis. Madison **52**(55–66), 11 (2010)
14. Stahl, F., Schomm, F., Vossen, G., Vomfell, L.: A classification framework for data marketplaces. Vietnam J. Comput. Sci. pp. 1–7 (2016)
15. Tran, T., Nguyen, D.T.: Algorithm of computing verbal relationships for generating vietnamese paragraph of summarization from the logical expression of discourse representation structure. Vietnam J. Comput. Sci. pp. 1–12 (2015)
16. Tran, V.C., Hwang, D., Jung, J.J.: Semi-supervised approach based on co-occurrence coefficient for named entity recognition on twitter. In: 2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS), pp. 141–146. IEEE (2015)
17. Yao, L., Sun, C., Wang, X., Wang, X.: Combining self learning and active learning for chinese named entity recognition. J. Softw. **5**(5), 530–537 (2010)

Part V
Internet and Network Technologies

Prediction of Topics Popularity on On-Line Social Networks

František Babič and Anna Drábiková

Abstract Social networks represent nowadays an important communication channel for various groups of people over the world. They offer an audience for various activities with the aim to get the attention of the related target group, e.g. marketing or political campaigns. The aim of this paper is to understand the hid-den trends and various developments in such type of data and extract possible new and interesting knowledge for business purposes. For this purpose, we used data from two different social networks: Twitter and Tom's Hardware. The completely analytical process was realised in line with CRISP-DM methodology; we selected the suitable methods of machine learning and exploratory data analysis to get the expected results. The created decision support application offers a group of methods to understand the data within the exploratory analysis, to generate a prediction model with the highest accuracy or to extract the rules supporting decision process during an on-line campaign. The best-achieved accuracy was higher than 95 % and extracted rules represent a good basis to ensure an expected popularity for selected topics in the future. Although we tested the system within a dataset closer oriented to the ICT sector, we will evaluate its applicability on a wider scale in our future work.

Keywords Social network · Buzz · Analysis

1 Introduction

Social networks create a virtual space for different topics, contributors and com-munities. They are an important channel through which it is possible to test the potential commercial success of a new product or to test the potential popularity of

F. Babič (✉) · A. Drábiková
Faculty of Electrical Engineering and Informatics, Department of Cybernetics and Artificial Intelligence, Technical University of Kosice, Kosice, Slovakia
e-mail: frantisek.babic@tuke.sk

A. Drábiková
e-mail: anna.drabikova@tuke.sk

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_29

a given political party or a particular policy. This medium represents an important tool to make the marketing campaigns more effective and tailored. One of the most social phenomena on social networks is a buzz. Several topics in user's posts may cause an increased interest by other contributors and a high occurrence of new discussion threads in a relatively short time. Some experts point out that the buzz events can occur gradually or suddenly [12]. To our knowledge, the phenomenon of buzz represents an interesting research question because the existing works have not formally defined yet. The typical model of the buzz can include some rules containing the key descriptors as e.g. a topic, a level of the attention and a period of the attention. From the research point of view, it is important to find the most characteristic combinations of the available descriptors for the related buzz events. We can translate this objective into the language of the data mining as a classification task including a target binary class and a set of input variables representing the available descriptors of the topics on the investigated social network. The traditional combination of the exploratory data analysis and data from the social networks is a creation of the various visualisations of the extracted networks in a form of the graphs [15]. Our approach is oriented to the numerical descriptors with the aim to identify the possible hidden relations and time trends.

The content of this paper is organised as follows: a brief introduction with the carefully selected related works, a description of the performed analytical process in line with the CRISP-DM methodology, and a conclusion with a summary and the possible directions for the future work.

1.1 Related Work

Analysis of data from the social networks is becoming increasingly meaningful not only for providers of these networks themselves but also for the companies that offer their products and services or communicate with their customers through them. The effect of buzz events to predict the movements in financial markets describes the work of authors [20]. They used Twitter feeds collected for 5 months within four 2000-km wide circles with centres in Pittsburgh, Atlanta, Las Vegas and Boise respectively. Obtained results statistically show that public opinion measured from a large-scale collection of emotional re-tweets is correlated to the financial market movements and following four keywords were identified as the most important for such a prediction: dollar, $, gold, oil, job and economy. A collective of authors [11] used a neural network to predict the number of re-tweets on the Twitter. This approach combined following three features: popularity, expressivity and singularity for this prediction. The obtained result represents a 0.72 F1 score for the tweets that were forwarded at least 60 times. The authors [1] used more than 2 million of the Twitter tweets referring to the 24 different movies with the aim to discover a possible relationship between a popularity of the movie on the Twitter and its expected box office revenues. They constructed a linear regression model that outperformed in accuracy the Hollywood Stock Exchange—an artificial on-line

market. The authors [6] analysed the same dataset as was used in this paper within modified J48 algorithm, neural network and simple regression method available within Weka data mining software. The proposed modification resulted in more accurate classification models (MAENN = 251.85, MAEJ48 m = 80.23). The authors in [8] design a new tool to explore the raw data from the Twitter network. They use the interactive plots of all the data to examine the entire dataset for initial hypotheses to test, e.g. a heat map for the tweet location. The similar prototype of the system for a visual interactive analysis of large geo-referenced microblog datasets describes the work [18]. Also, this type of data can be processed by the methods from the area of Formal Concept Analysis, especially generalized one-sided concept lattices (theoretically introduced in [4]) which can be applied to various data tables and attribute types [3], and also computed in distributed environment for larger input datasets [5].

Finally, we can say that the analysis of the data from the social network provides an important source of knowledge for various purposes. Existing works are oriented to the design and implementation of the different tools offering mainly the visualisation techniques. Therefore, a new simple decision support application can be an interesting alternative how to understand the topics popularity.

2 Analytical Process

We performed the analytical process in accordance with the CRISP-DM methodology [7, 19] to ensure the correctly and clearly understandable sequence of applied operations and methods. This methodology defines six main phases:

- Business understanding is oriented to specification of business goal, followed with transformation of specified business goal to concrete model-ling task(s). Based on this specification, relevant mining methods and algorithms are selected with necessary resources.
- Data understanding covers collection of necessary input data for specified tasks; its understanding and initial description with basic statistical characteristics.
- Data preparation is usually the most complex and also most time consuming phase of the whole data mining process (usually taking 60–70 % of the overall time). The goal of relevant data operations is to prepare final version of a dataset based on requirements of selected algorithm that will be applied in the next phase.
- Modelling—deals with an application of suitable data mining algorithms on the pre-processed data.
- Evaluation—phase is oriented towards the evaluation of generated models and obtained results based on specified goals in business understanding.
- Deployment contains the exploitation of created mining models in real cases, their adaptation, maintenance and collection of acquired experiences and knowledge.

2.1 Business Understanding

Buzz events represent important information for an organisation of the marketing or political campaigns through social media. If we could identify the applicable rules for this phenomenon, we will be able to make the new product/topic promotion process more effective. The more effective process generates more customers or supporters that mean higher profit. An important factor, in this case, is the potential dependence of the extracted rules on the relevant domain, so we oriented our experiments on the IT with relevant topics such as graphic cards, overclocking, OS Android, etc. In terms of knowledge discovery, it is possible to transfer presented business goal into following mining goal: binary classification to predict if related event will be buzz (1) or not (0). For this purpose, we selected several algorithms suitable for generation of decision trees models from which it was possible to extract some interesting decision rules and hidden relations between related attributes.

A decision tree is a flowchart-like tree structure, where each non-leaf node represents a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent target classes or class distributions [13]. The beginnings in this domain represent algorithms as ID3 [16], C4.5 [17], C5.0 or CART [2]. Both C4.5 and C5.0 generate the decision models based on tress or rules set, but the second one is faster, requires a less memory, generates smaller decision trees and has a greater ability to process classes that have very low representation in the training data [10]. CART (Classification And Regression Trees) implementation is very similar to C4.5; only the CART constructs the tree based on a numerical splitting criterion recursively applied to the data. Resulting binary decision trees are more sparing with data and detect more patterns/rules/structures before too little data are left for learning. CART uses GINI Index to determine in which attribute the branch should be generated [14].

2.2 Data Understanding and Preparation

Data used in this article were published by data analysis, modelling and machine learning group (AMA), Grenoble Computer Science Laboratory [9]. It represents an activity of the users on two different social networks (Twitter and Tom's Hardware) within 6671 topics from IT domain. Twitter (TW) is an online social networking service that enables users to send and read short 140-character mes-sages called "tweets" (Wikipedia). Tom's Hardware (TH) is a worldwide forum providing articles, news, price comparisons, videos and reviews on computer hardware and high technology (Wikipedia).

Initial TW dataset contained more than 140 thousand of records described by 77 attributes; in the case of TH's dataset, more than 7.9 thousand of records are available described by 96 attributes. Both datasets hadn't any missing values, but

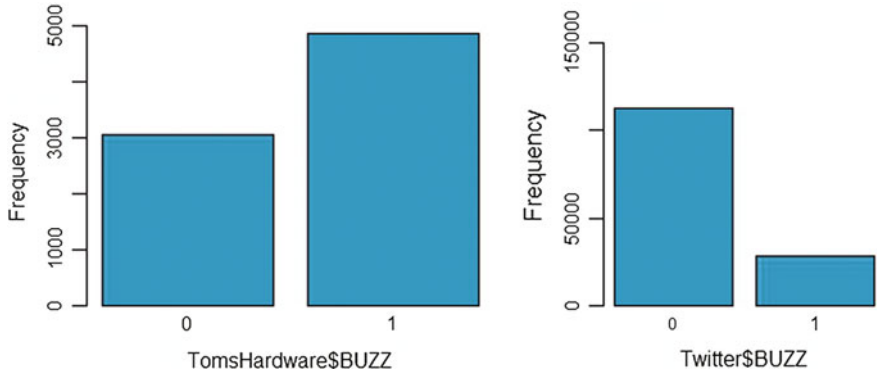


Fig. 1 Distribution of the target attribute's values in both datasets (*x-axis* target classes, *y-axis* number of records for the related class/frequency)

they were constructed within the different size of time windows: TW with 7-days (t0, t1, …, t6) and TH with 8-weeks (t0, t1, …, t7). Following histograms visualise the initial distribution of the target attribute for both datasets (Fig. 1).

The members of the group constructed a binary level of the target attribute within two different conditions:

- If the level of topic's popularity in a given time window reaches or exceeds a certain threshold (500), the topic is considered a buzz.
- If the level of topic's popularity in given time window increases by the certain threshold (500), the topic is a buzz.

It means that we had available several datasets with the different labelling of the target attribute (we used the first one for our experiments), but a similar list of in-put variables described the each record (topic) and these variables changed in the relevant time window, see Table 1.

At first, we performed a basic statistical understanding of both datasets, i.e. we calculated a minimum, maximum and mean value for every included numeric at-tribute. Next, we decided to visualise and compare the value's development for relevant attributes at a time in order to identify potential differences.

We identified the main differences in TH's dataset within the attributes NCD, NAD, AI, NAC and NA. This is the first finding, which attributes can have an influence on the fact that related topic will be buzz event or not.

Analysing this set of graphs further, it is possible to identify the points in time in which the trend was a change, e.g. in the case of attribute AI, the number of new authors interacting on the instance's topic at time t have increased significantly until 4th time point, after which this value was constant or decreased (buzz = 1). We can observer a similar situation within attribute NA (number of authors interacting on the instance's topic at time t).

Similar visual exploration was performed within TW dataset and findings were little different as in the case of TH, e.g. time development for attribute NCD was

Table 1 Attributes description for data understanding

Name	Description
Number of created discussions (NCD)	A number of discussions created at time step t and involving the instance's topic
Attention level (AS_NA)	A measure of the attention paid to the instance's topic on a social network (measured with a number of authors)
Author increase (AI)	A number of new authors interacting on the instance's topic at time t
Burstiness level (BL)	The ratio between attributes NCD and NAD for the relevant topic at time t
Number of discussions (NAD)	A number of discussions involving the instance's topic until time t
Average discussions length (ADL)	The average length of a discussion belonging to the instance's topic
Number of displays (ND)	A number of times discussions relying on the instance's topic have been displayed by users *It is not included in Twitter dataset*
Number of atomic Containers (NAC)	A total number of atomic containers generated through the whole social network on the instance's topic until time t
Attention level (AS_NAC)	A measure of the attention paid to the instance's topic on a social network (measured with a number of contributions)
Contribution sparseness (CS)	A measure of spreading of contributions over discussion for the instance's topic at time t
Author interaction (AT)	The average number of authors interacting on the instance's topic within a discussion
Number of authors (NA)	A number of authors interacting on the instance's topic at time t

very similar in both cases of target buzz values. We identified the main difference within attributes AT and ADL, see following Fig. 2.

2.3 *Modelling and Evaluation*

The C4.5, C5.0 and CART algorithms application resulted in the number of experiments, which have been oriented on a generation of the most accurate predictive models and potentially interesting rules. We compared the generated models using traditional matrix (see Table 2) representing the ability of the relevant models to classify correctly both values of the target attribute. This matrix contains the following variables: TP (true positive)—correctly classified buzz = 0, FP (false positive)—incorrectly classified buzz = 1, FN (false negative)—incorrectly classified buzz = 0, TN (true negative)—correctly classified buzz = 1 (Table 3).

Table 3 presents the best results generated within TH dataset. We can conclude that obtained accuracy is high and similar in all three cases. We obtained the similar results within a different data division between the training and testing set (80/20).

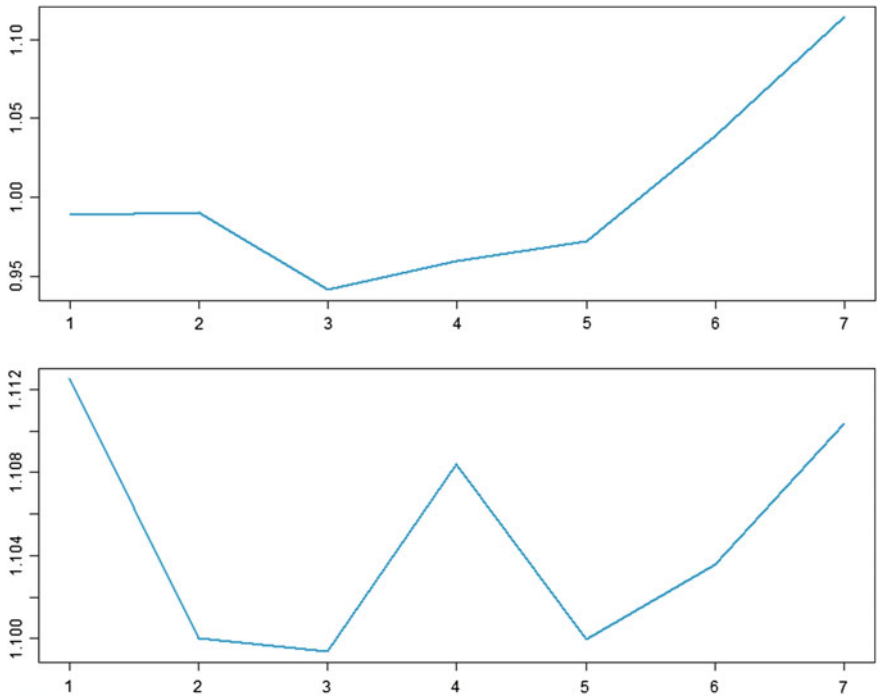


Fig. 2 Comparison of the time developments for the AT attribute in the TW data (*top* buzz = 0, *lower* buzz = 1, *x-axis* calculated average values, *y-axis* time window)

Table 2 Attributes description for data understanding

		True values	
		0	1
Predicted values	0	TP	FP
	1	FN	TN

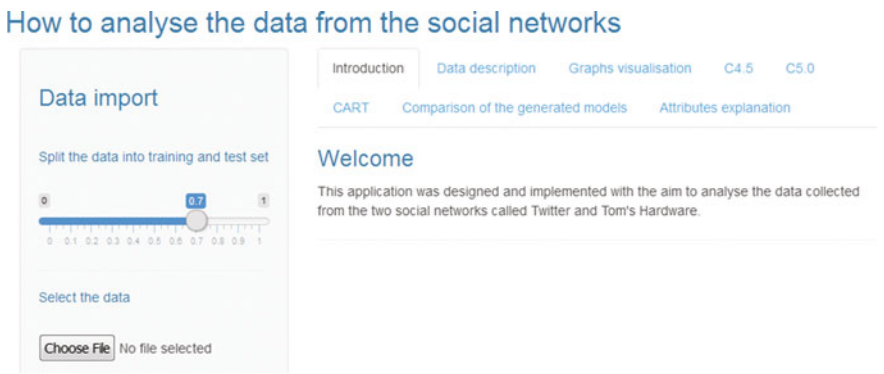


Fig. 3 Example of the created simple decision support application

Table 3 Models with the highest accuracy generated by the 3 selected algorithms (TH dataset)

	TH dataset (70/30)				
	TP	FP	FN	TN	ACC (%)
C4.5	894	28	39	1 411	97.18
C5.0	889	33	28	1 422	97.42
CART	880	42	25	1 425	97.18

Table 4 Models with the highest accuracy generated by the 3 selected algorithms (TW dataset)

	TW dataset (70/30)				
	TP	FP	FN	TN	ACC (%)
C4.5	33 290	632	807	7 484	96.59
C5.0	33 336	586	589	7 432	96.57
CART	33 373	549	1037	7 254	96.24

Following table (Table 4) presents the most accurate models generated over TW dataset. For an evaluation of the prediction quality, it is necessary to consider how many records were classified incorrectly (FP, FN). There are two types of the errors: a model predicts the record as a buzz, but it does not reach the expected level of popularity; or the other way around. In our case, an improvement was oriented to the FN's lower value.

The final buzz model for both datasets includes not only the descriptors related to the number of the authors interacting on the instance's topic at time t and related increase; or an overall number of created discussions and their continuous increase. In addition, the rules containing these attributes ordered in time (the first three rules are relevant for the TH, the last for the TW):

- IF the "number of displayed discussions relevant to the instance's topic" in the 2nd week was >230 AND the "number of displayed discussions relevant to the instance's topic" in the 8th week was >270 THEN this topic becomes a buzz.
- IF the "number of displayed discussions relevant to the instance's topic" was in the 2nd week $\leq$230 AND the "average number of author's interacting on the instance's topic within a discussion" in the 4th week was $\leq$5 AND the "number of displayed discussions relevant to the instance's topic" in the 8th week was >270 AND then this topic becomes a buzz.
- IF the "number of displayed discussions relevant to the topic" in the 2nd week was $\leq$230 AND the "number of displayed discussions relevant to the topic" was in the 5th week $\leq$245 AND "number of authors interacting on the instance's topic" in 5th week was $\leq$2 AND "attention level" in the 7th week was $\leq$0.0023 AND "number of displayed discussions relevant to the topic" was in the 8th week >270 THEN this topic becomes a buzz.
- IF the "number of created discussions involved the instance's topic" on the 1st day >255 the "total number of atomic containers generated through the whole Twitter social network on the instance's topic" on the 2nd day $\leq$255 AND "total number of atomic containers generated through the whole Twitter social network on the instance's topic" on the 6th day >281 THEN this topic becomes a buzz.

2.4 Deployment

This final phase of the whole CRISP-DM analytical cycle deals with different approaches on how to use the generated models or extracted knowledge in real cases. In our case, we decided to design and implemented a supporting decision application for buzz predictions.

We implemented this application in R—language that represents a free software environment for statistical computing and graphics. Available prototype within shinyapps.io provides several features necessary to import, understand and analyse different data from social networks related to the target task—binary classification of the buzz events.

3 Conclusion

Analysis of user behaviour on social networks represents an important source of information further using for various purposes. Various companies use the social media for an organisation of their marketing campaigns or the politicians to influence the opinion of their voters. The motivation is to reach a level of a popularity, which will give us the expected attention. In addition, we can describe this level with various descriptors or related rules, both extracted from the historical data representing different time trend on the social media. We aimed at the two avail-able datasets Twitter and Tom's Hardware to investigate potential hidden knowledge and time trends for the topics popularity. We selected two possible approaches to meet this goal: an exploratory data analysis and a binary classification. The first one included some visualisations in a form of graphs, e.g. a histogram or a run chart with the aim to find the hidden differences with a strong influence on the possible popularity of the relevant topic in the future. We calculated and visualised the time trends for all attributes especially for the buzz and no-buzz topics (Fig. 2 as an example). The second direction resulted in a number of generated classification models evaluated by the typical matrix. The obtained results were plausible; we achieved the accuracy higher that 95 %. In addition, the models with the highest accuracy form the basis to extract the decision rules. Finally, we integrated the all performed operations into one simple application implemented within R language. This application can be used for an understanding of the collected data representing the popularity's development of the topics on the social networks. The future work will focus on the improvement of the available version and on the analysis of the data with the relative labelling.

Acknowledgments The work presented in this paper was partially supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/0493/16, by the Cultural and Educational Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 025TUKE-4/2015 and by the internal faculty research project no. FEI-2015-2.

References

1. Asur, S., Huberman, B.A.: Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 492–99 (2010)
2. Breiman, L.: Classification and Regression Trees. Repr. Chapman & Hall, Boca Raton (1998)
3. Butka, P., Pócs, J., Pócsová, J., Sarnovský, M.: Multiple data tables processing via one-sided concept lattices. Adv. Intell. Syst. Comput. **183**, 89–98 (2013)
4. Butka, P., Pócs, J., Pócsová, J.: On equivalence of conceptual scaling and generalized one-sided concept lattices. Inf. Sci. **259**, 57–70 (2014)
5. Butka, P., Pócs, J., Pócsová, J.: Distributed computation of generalized one-sided concept lattices on sparse data tables. Comput. Inform. **34**(1), 77–98 (2015)
6. Dahiya, J., Rahsmi, K.: Prediction of popularity on social web. Int. J. Comput. Sci. Manage. Stud. **14**(7), 37–43 (2014)
7. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 Step-by-Step Data Mining Guide (2000)
8. Cheng, D., Schretlen, P., Wright, W.: Tile based visual analytics for twitter big data exploratory analysis. In: Proceedings of 2013 IEEE International Conference on Big Data, Santa Clara, USA, pp. 2–4 (2013)
9. Kawala, F., Douzal-Chouakria, A., Gaussier, E., Dimert, E.: Prédictions D'activité Dans Les Réseaux Sociaux En Ligne. In: Actes de La Conférence Sur Les Modèles et l′Analyse Des Réseaux: Approches Mathématiques et Informatique (2013)
10. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer, New York (2013)
11. Morchid, M., Linarès, G., Dufour, R..: Characterizing and predicting bursty events: the buzz case study on twitter. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC, Reykjavik, Iceland, pp. 2766–2771 (2014)
12. Mourdoukoutas, P., Siomkos, G.J.: The Seven Principles of WOM and Buzz Marketing, Springer (2010)
13. Murthy, S.K.: Automatic construction of decision trees from data: a multi-disciplinary survey. Data Min. Knowl. Disc. **2**(4), 345–389 (1998)
14. Patil, N., Rekha, L., Vidya Ch.: Comparison of C5.0 and cart classification algorithms using pruning technique. Int. J. Eng. Res. Technol. **1**(4), 1–5 (2012)
15. Peter, A., Shneiderman, B.: Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, pp. 265–274 (2008)
16. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
17. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993)
18. Schreck, T., Keim, D.: Visual analysis of social media data. Computer **46**(5), 68–75 (2013)
19. Shearer, C.: The CRISP-DM model: the new blueprint for data mining. J. Data Ware-Housing **5**(4), 13–22 (2000)
20. Zhang, X., Fuehres, H., Gloor, P.A.: Predicting asset value through twitter buzz. In: Altmann, J, Baumöl, U., Krämer, B.J. (eds.) Advances in Collective Intelligence 2011, vol. 113, pp. 23–34. Berlin, Heidelberg, Springer (2012)

eanaliza.pl—A New Online Service for Financial Analysis

Tomasz Jastrząb, Monika Wieczorek-Kosmala, Joanna Błach and Grzegorz Kwiatkowski

Abstract The paper presents a new online service, which provides an extensive support for financial analysis. The service is targeted at various groups of end-users, including small companies and their stakeholders, accountancy offices as well as financial advisors. The scope of the analysis involves five groups of financial ratios as well as percentage structure and dynamics analysis. The aim of the paper is to show the interdisciplinary and multidimensional nature of the service viewed from several different perspectives. The paper also aims at presenting certain programmatic solutions that we found useful and interesting while implementing this financial service.

Keywords Online services · Financial analysis · e-Finance · e-Business · e-Commerce

1 Introduction

Financial analysis is a crucial element of effective financial management. It allows to transform financial data acquired from the accounting system into a set of financial ratios, useful in the decision-making process. The role and importance of financial

T. Jastrząb (✉) · G. Kwiatkowski
Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland
e-mail: Tomasz.Jastrzab@polsl.pl

G. Kwiatkowski
e-mail: Grzegorz.Wojciech.Kwiatkowski@polsl.pl

T. Jastrząb
Technicenter Sp. z o.o, ul. Tatrzańska 2, 41-907 Bytom, Poland

M. Wieczorek-Kosmala · J. Błach
University of Economics, ul. Bogucicka 14, 40-226 Katowice, Poland
e-mail: monika.wieczorek-kosmala@ue.katowice.pl

J. Błach
e-mail: jblach@ue.katowice.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_30

analysis can be considered both from the company's perspective (its managers as internal users of financial analysis) and its stakeholders (as external users of financial analysis). Thus, internal financial analysis, by providing the opportunity to construct the intra-company system of financial ratios, together with clearly described framework of their interpretation, may strongly enhance the efficiency of the company. As for external users, financial analysis is important, since it provides the complex analytical approach applicable by existing and potential investors, creditors, customers, suppliers and other groups of stakeholders.

The main contribution of the paper is the presentation of an innovative online service[1] which provides the ability to perform an extensive financial analysis. The key features of the service, apart from its economic importance, include: dedicated functionalities targeted at different groups of end-users, a hierarchy of access rights and user roles aiding the company's cooperation with other parties, easy integration with selected existing accountancy tools as well as thorough e-Learning capabilities. According to our best knowledge there are no other online services providing comparable functionality. The existing solutions that also deal with financial analysis are usually Excel-based and tend to be designed to fulfill the demands of companies ordering the software. Our service aims to support much broader range of potential customers.

The paper is divided into five sections. Section 2 discusses briefly the scope of the analysis provided by the service. Section 3 presents the service from different perspectives. Section 4 describes the main workflow. Section 5 reports some results of system testing. Finally, Sect. 6 contains the summary and future perspectives.

2 The Economic Background

As a managerial tool, financial analysis offers a complex set of financial ratios applicable in the assessment of company's performance, in particular its efficiency and financial stability. In the retrospective context, financial analysis is applicable for the assessment of company's past performance, whereas in the prospective context it is applicable for forecasting company's future situation and for supporting the settlement of strategic goals. It is also used to examine risk and expected returns. The construction of financial ratios allows to do comparisons between the assessed companies (cross-sectional analysis) or against the given benchmarks (e.g. industry average), as well as to do time-comparisons in the dynamic context (trend analysis) [1].

Typically, financial analysis is used to evaluate five aspects of operating performance and financial condition of a company [2]. Each of the five aspects is illustrated by a given set of financial ratios based on the financial statement data (mostly: balance sheet, profit and loss account and cash flow statement). One shall remember

[1] The web page of the service and the service itself can be found under https://eanaliza.pl.

that the interpretation of the results reflecting one of these aspects often depends on the results obtained for the other as they are highly intercorrelated [3]. However, the distinction of the analytical fields is relatively hermetic as regards a defined managerial context. Usually, five aspects of company's financial situation form the subject of financial analysis: (1) general performance, (2) financial liquidity, (3) debt management, (4) efficiency (activity) and (5) profitability [4].

General performance analysis is designed to give a preliminary outlook on the company's financial stability. It shall open the analysis, as it allows to identify whether the structure of funds (capital structure) is adjusted to the structure of assets [4]. In the case of wrong adjustment, the synthetic ratio of financial situation is decreasing, which indicates the worsening of a company's financial stability. The interpretation of other groups of financial ratios explains the reasons for the observed changes of financial stability.

Financial liquidity analysis supports one of the most relevant short-term aspects of financial management. It allows to verify whether the company's liquid assets (cash and marketable securities) are high enough to cover company's short-term obligations. Any deficiencies of liquid assets may in turn lead to bankruptcy, thus the maintenance of financial liquidity is critical. Financial liquidity analysis is based on the application of financial ratios which confront the volume of liquid (current) assets with the volume of company's current liabilities e.g. current ratio, quick ratio [5]. In the extended view, the analysis refers also to the level of net working capital (which reflects the surplus of long-term funds), relative to company's assets. The assessment of financial liquidity may be also performed in the dynamic context, by the application of financial ratios based on company's cash flows [6].

Debt management analysis refers to the capital structure of the company. It revises the structure of funds (the extent to which a company uses debt financing) and additionally verifies the company's ability to pay interest and to service debt (interests and principal repayment) [7]. Debt management ratios can alert management to the need for changing the company's financing mix before insolvency problems develop [8].

Efficiency analysis (also known as activity analysis) allows to verify how effectively the company uses its assets to generate sales revenues, reflected in a group of asset productivity (asset management) ratios [9]. The efficiency is also evaluated from the company's operating performance point of view, by examining the proportions between costs and revenues. Altogether efficiency analysis consists of: turnover ratios, cash conversion cycle and level of costs.

Profitability analysis is designed to verify the company's ability to generate profits on the performed activity. From the profit and loss account point of view, profitability analysis addresses the ability to generate profits (rate on return) on assets, sales and equity [5]. Profitability ratios show the combined effects of liquidity, asset management and debt on operating results [9].

From the end-users perspective, financial analysis performs a dual role. For the internal end-users (the company's management) it offers a vital tool of effective managerial decisions, both in the retrospective and prospective context. By monitoring changes in financial ratios, managers can spot developing areas of strength and weaknesses and can take appropriate action to modify the company's development

strategy. However, financial analysis is also applicable for the needs of various groups of external end-users (company's stakeholders), as it relies mostly on the data presented in the financial statements. In particular, the creditors and shareholders may implement financial ratios to conduct the complex analysis of financial performance of a company subject to credit decisions (as creditworthiness assessment) or investment decision (as a part of fundamental analysis). However, the successful implementation of financial analysis relies on the proper understanding of financial ratios and the information provided by their level or direction of the changes over time. It is also worth stressing that ratios analysis is subject to several limitations that should be taken into account while assessing and comparing company's performance [10, 11].

It seems that there is a need for complex analytical tools, addressing the needs of both internal and external end-users of financial analysis and enhancing the proper understanding and application of financial ratios.

3 Overview of the System

Due to the interdisciplinary nature of the discussed service, which combines the domains of IT, computer science and economics, the importance and usefulness of the service can be viewed from different perspectives discussed in the following sections.

3.1 e-Business/e-Commerce

Before we present the relation of eanaliza.pl to the areas of e-Business and e-Commerce let us clarify what will be understood by these terms. It can be observed that precise definitions of e-Business and e-Commerce are not well established in the literature, see for example [12–14]. Coltman et. al [15] propose to define e-Business as any business conducted using digital infrastructure, which is also similar to the definition proposed in [16]. Kalakota and Robinson [17] present another point of view on the relation between e-Business and e-Commerce, namely that e-Business is the third phase of the evolution of e-Commerce, in which the Internet affects profitability. In the sequel, following the considerations in [17] we will consider that e-Business and e-Commerce represent:

- online accessibility,
- online business-to-consumer (B2C), business-to-business (B2B) and intra-business relations,
- online documents management.

The first aspect of e-Business/e-Commerce namely online accessibility is realized thanks to the utilized Software as a Service (SaaS) distribution model. It is mainly

reflected in the fact that the use of the service does not require any client-side software installation or configuration. Furthermore, users of the service are provided with online support, both through the web page as well as the service itself, which further increases its user-friendliness and accessibility. What is more, the integration of the service with existing accountancy tools such as Comarch Optima ERP or Sage Symfonia makes the service available for a wide range of customers. Finally, the service is targeted at different groups of end-users, namely small companies and their stakeholders, accountancy offices and financial advisory offices. As a consequence, it should be viewed as a multi-purpose tool, e.g. for individuals it may help in monitoring company's finances, for accountancy offices it may be considered as an added-value to their basic scope of service, whereas for financial advisors it could help to offer the best-suited financial products.

The second element of e-Business/e-Commerce has been implemented in the system with the use of access rights, user roles and system objects representing contractors.[2] In particular, the intra-business relations between employees and CEOs as well as between different employees are modeled by *owner* and *user* access roles. They correspond directly to different privileges reflected in access restrictions to certain views or limitations of certain functionalities, e.g. only the owner is able to configure the information about company's account or access payment details. On the other hand, the B2B and B2C relations are managed through the *observer* role. An observer is a user of the system with access rights limited to: personal data configuration and analysis results preview (for the set of assigned contractors). In terms of B2B relations one can imagine that an observer can be the company for which the financial advisor or accountancy office are providing their service. On the other hand, B2C relations may involve giving the observer access right to stock and/or shareholders, who might be interested in monitoring the company's overall condition.

The third issue related to e-Business/e-Commerce, which is online documents management, is reflected in three elements of eanaliza.pl. The first element is the use of electronic documents as data sources for the financial analysis process (see Sect. 4). The second element is the availability of analysis results in the form of downloadable PDF or CSV files. The latter format is intended to be usable for systems performing further processing, while PDF files are to be used for reporting purposes. The last element regards electronic invoices, which are issued as a confirmation of the payments made.

3.2 e-Finance

eanaliza.pl is certainly a system lying in the domain of e-Finance, since it provides financial services using electronic communication and computation [18]. It provides a broad scope of financial analysis, as discussed in Sect. 2 and may be useful for

[2] A contractor is considered to be a client of a customer of eanaliza.pl service.

various financial parties and institutions (for examples see Sect. 3.1). Moreover it provides some unique features that can be of real value to accountants, financial advisors and experts. Among these features one may find, e.g. *the error detection system* discussed in detail in Sect. 4, which may help an accountant to verify the correctness of company's financial documents, or *additional* and *user-defined ratios*, which can be used for monitoring the aspects of company's finances falling outside of the scope of basic financial analysis. Thanks to the flexible design of the ratio evaluation process the definition of own ratios is straightforward. To define a new ratio it is enough to specify its name, formula and a group of ratios to which it should belong. However, one may also specify an extensive description of the ratio, which will be later displayed in a tooltip.

3.3 e-Learning

eanaliza.pl can also be viewed as an e-Learning system thanks to the extensive and informative web page content as well as the elements of the service itself. In particular, the service provides two essential elements increasing its user-friendliness and e-Learning capabilities. Firstly, it provides the user with an exhaustive information relevant to the currently viewed part of the system (*context help*). Secondly, to increase user awareness and improve knowledge on the elements of the financial analysis a set of extensive tooltips is provided.

The design of the context help allows an experienced user to hide the unneeded information, while preserving the ability to bring the help back whenever required. The aim of the tooltips is to present a detailed information on the way particular ratios are computed as well as what is their meaning regarding the overall company's condition and how should the computed values be interpreted. Finally, we are currently working on two additional aspects that could make the system even more important for the e-Learning purposes. These elements are a multi-lingual version of the system, including English and German, and a student version available for free. The latter solution should be of particular interest, since it could be of help to both teachers and students.

4 The Basic Workflow

This section discusses the key element of the service, namely the ability to analyze and monitor company's finances. The financial analysis workflow provided by the service is a six-step process that can be briefly summarized by the following points:

- input data provision,
- input data configuration,
- input data verification,

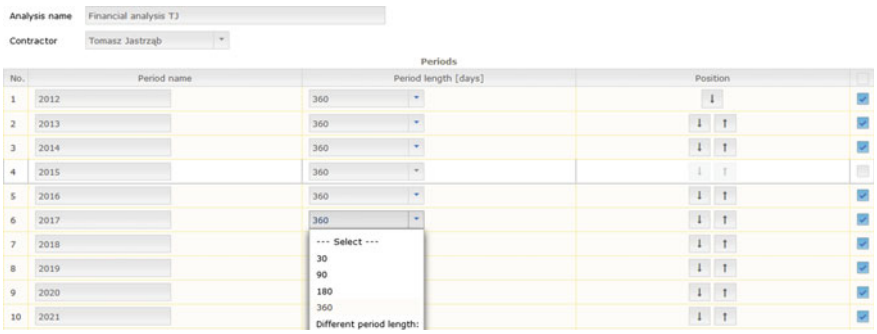


Fig. 1 Input data configuration step, including: analysis name (*top*), contractor (*center*) and periods (*bottom*) specification

- additional data provision,
- data processing,
- results visualization and interpretation.

Let us now elaborate on each of the above mentioned elements of the workflow. As already mentioned in Sect. 2 there are two main components that have to be supplied to perform the financial analysis, namely *the balance sheet* and *the profit and loss account*. In eanaliza.pl these elements can be provided in one of three ways. The system supports: manual data provision, semi-automatic data provision with Excel spreadsheets and semi-automatic data provision with PDF files. Both Excel spreadsheets and PDF files can be generated from existing software tools supporting accountancy provided that they comply with [19]. In particular, the profit and loss account should follow the rules of the comparative version of the statement as given in [19]. Additionally in case of Excel spreadsheets the system provides ready-to-use templates available for download. In case of semi-automatic data provision some basic verification of data correctness is performed. Namely, it is required that the number of data periods included in both files (i.e. balance sheet and profit and loss account) is the same and not smaller than two.

The step of input data configuration involves three levels of configuration. Firstly, the user is supposed to provide a name of the newly created analysis (cf. the topmost part of Fig. 1). Secondly, the analysis has to be assigned to one of available contractors (cf. the middle part of Fig. 1). This step is necessary to facilitate the process of access rights management, especially for observers, as discussed in Sect. 3.1. Finally, the most crucial part of the configuration step is the definition of data periods included in the analysis (cf. bottom part of Fig. 1). There are four elements of data periods definition which can be changed, depicted left-to-right in the table in Fig. 1. Firstly, the names of the periods can be modified.[3] Secondly, the length (in days) of each period should be specified—it can be either selected from the predefined list or provided manually if the desired length is not available (see the dropdown

[3]Note that the names are either extracted from the files or some defaults are automatically supplied.

list in Fig. 1). Thirdly, the mutual order of periods can be modified, i.e. periods can be moved before or after some other periods. Finally, the range of periods can be restricted by selection or deselection of some of them.

The third step of the workflow involves the verification of balance sheet and profit and loss account data. Its main purpose is to preview the data read from the input files. The step can be also used to correct some mistakenly read values, e.g. if the order of input data rows was not fully compliant with the expected data template. For the manual data provision mode, this step allows for the input of the data.

The fourth step allows to provide the data that lies outside of the scope of the balance sheet and profit and loss account but is required to perform the complete financial analysis. Currently, the additional elements involve only the information on short- and long-term loans, which indirectly influence certain analytical ratios, such as e.g. modified current ratio, modified quick ratio or cash flow coverage ratio. The indirect influence results from the fact that, among others the aforementioned ratios, include in their definitions the current principal repayment. The current principal repayment, in turn, can be precisely computed only if the additional data on short- and long-term loans is provided.

The data processing step is the core part of the workflow in which all the ratios are actually computed basing on the data provided in the previous steps. However, to make the solution flexible and easy to extend in the future, instead of incorporating all the calculations directly in the code, we decided to define each ratio with a string representing its formula. Then to evaluate each ratio it is enough to compute the Reverse Polish Notation form of the formula and evaluate the resulting expression. For the purpose of reversal and evaluation of the formulae we have used the algorithms described in [20, 21].

The final step, i.e. results visualization and interpretation can take one of two forms. If in the course of data processing some errors in the input data have been detected, the computed results will not be made visible until the errors are accepted or corrected. On the other hand, given correct input data, the user will be able to view the results in graphical and tabular form as well as to interpret or discuss their economic meaning and relevance. To give a better understanding of the possible outcomes of the analysis let us consider two cases depicted in Figs. 2 and 3.

Figure 2 presents the outcome of the situation in which some errors have been detected in the data processing phase. However, it should be clarified that the nature of the errors is purely economic and not related to the service itself. As already mentioned, an error can be either *accepted* or *corrected*. The former action should be understood as a decision to ignore the error. Errors that can be ignored may be related to the following problems:

- the input data is severely corrupted. In this case there is no easy way, other than the provision of new input data, to correct the error. An example of such an error is that the totals of assets and liabilities in the balance sheet do not match.
- the input data is in fact correct, but only taking into account certain special circumstances. An example of this kind of error would be a zero amortisation and depreciation value and a positive sum of intangible and tangible fixed assets. This can

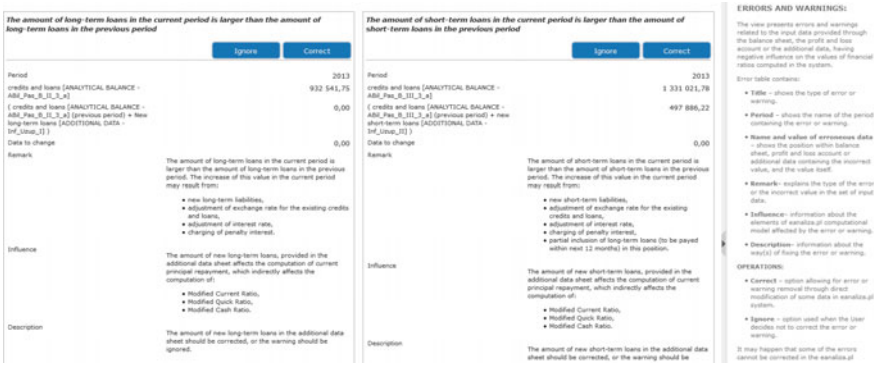


Fig. 2 Results view for erroneous input data, including: error descriptions (*left*, *center*) and context help (*right*)

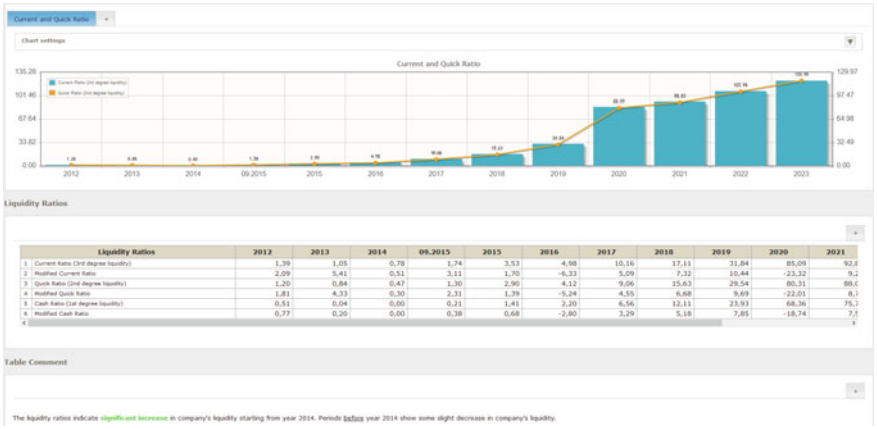


Fig. 3 Results view for correct input data, including: graphical (*top*) and tabular (*center*) form of data presentation and a comment (*bottom*)

happen for instance if the assets include lands which are not in perpetual usufruct and thus are not amortised.

On the other hand, errors that can be corrected result from inconsistencies in the short- and long-term loans and can be easily amended using the provision (or modification) of additional data. The change in the value of either loan type for any data period triggers the recomputation of the ratios. This in turn may result in the appearance of some new errors, if the data was modified improperly.

From the implementation point of view, the error detection mechanism was implemented as a rule-based system having the total of 108 if-then rules specifying the conditions for erroneous situations. The conditions are represented by formulae composed of arithmetic and logical operators as well as digits and input data

elements (described by symbols). Rules describing two of the abovementioned situations could then be defined as follows:

- zero amortisation and depreciation (PlsAd) with positive sum of intangible assets (BalAssIa) and tangible fixed assets (BalAssTfa)

$$\textbf{if } \mathrm{PlsAd} = 0 \,\&\&\, (\mathrm{BalAssIa} + \mathrm{BalAssTfa}) > 0 \textbf{ then } \mathrm{error}\,, \tag{1}$$

- larger amount of long-term loans in the current period (CUR(BalLiaLtl)) than in the previous one (PREV(BalLiaLtl), CUR(AddLtl))

$$\textbf{if } \mathrm{CUR}(\mathrm{BalLiaLtl}) > (\mathrm{PREV}(\mathrm{BalLiaLtl}) + \mathrm{CUR}(\mathrm{AddLtl})) \textbf{ then } \mathrm{error}\,, \tag{2}$$

  where CUR and PREV denote the current and previous period, respectively and AddLtl comes from the additional data provision step.

Figure 3 presents the exemplary results obtained for a correctly prepared set of input data. The topmost part of Fig. 3 contains charts of selected ratios. It is possible to create up to five different charts with at most three data series each, separately for each group of ratios (giving the total of up to 90 charts). The charts can be also further configured as to the legend positioning, chart type or data tips inclusion. To make the use of eanaliza.pl easier a three-level hierarchy of chart templates has been implemented, including:

- default template—this template contains several charts that are automatically provided to each new user of the service,
- user-wise template—service users can define their own templates, which are then automatically applied to each new analysis performed by the user,
- analysis-wise template—default and user-wise templates can be also modified on a per-analysis basis.

The central part of Fig. 3 contains tabular version of the calculated results to aid an in-depth analysis of obtained results. This section allows also to modify the set of visible ratios by showing some additional ones. However, the basic set of ratios is always visible. Similarly to the case of charts it is possible to define user- and analysis-wise templates regarding the additional ratios visibility. Moreover, the tabular form plays also an important educational role, as mentioned in Sect. 3. By hovering the mouse pointer over any ratio, the user can display a thorough description of the ratio, including: a formula, a short discussion of the ratio's meaning and an interpretation of its possible values.

Finally, the bottom part is intended for expert users to discuss and interpret the results shown in graphical or tabular form. It allows to add some, possibly styled, comments to the results. These comments are also automatically included in the PDF report that can be later generated.

5 System Testing

Due to space limitations let us elaborate only on selected elements of eanaliza.pl system testing. In particular, let us discuss the automated unit testing of the error detection mechanism as well as some of the performed usability tests.

To verify the correctness of the calculations we have used the fact that both if-then rules in the error detection system, as well as formulae for financial ratios are stored in the system as strings, which can be easily evaluated after conversion to Reverse Polish Notation (see Sect. 4). Thus, these elements of the system could be tested automatically, without the need for interaction with the user interface. For this purpose we have prepared an extensive set of unit tests using the Java-based JUnit framework [22]. To get an estimate of the number of unit tests that were created let us consider that on average, each of the 108 if-then rules (see Sect. 4), required three different test cases (for the abnormal, border and normal situations), yielding as a consequence over 300 test cases for the error detection system alone. An interesting aspect of unit testing related to the error detection system was that the tests allowed to notice some mutual relations between various errors. For instance, while testing for negative value in the total of balance assets, it should be remembered that the total of balance liabilities has to be set accordingly, otherwise the error of mismatched balance totals will also be generated.

The usability evaluation involved the use of selected methods classified into *Testing* and *Inquiry* method classes by Ivory and Hearst [23]. As for the Testing method class we have selected Coaching Method and Codiscovery Learning, while in the Inquiry method class the Interviews and User Feedback approaches were taken. The first two usability evaluation methods were targeted at testing various scenarios of user role assignments and configurations. In particular, the elements that were tested included among others:

- the access to multiple accounts, possibly with different roles. This test aimed also at assessing the ease of switching between accessible accounts and being able to track the currently selected account,
- view and activity restrictions related to particular user roles. This test involved also the verification of the correct system behavior regarding the assignment of contractors to users and/or observers,
- the limits on the number of contractors, users and observers resulting from the selection of particular payment plans. This test allowed also to verify whether the scope of activities that are connected with each payment plan is suitable for the target users.

The use of Coaching Method, allowing the users to ask the expert (system developer) about certain functionalities, was mainly used in the initial stage of user roles testing. The questions were related e.g. to the possibility of having multiple owner accounts, or the order of gaining access to various accounts. As for the latter issue, the tests have shown that the order is insignificant, i.e. it is possible to both have an owner account and be given the user or observer access to some other account, as well

as to first be given a user or observer role and then to register as an owner of another account. A crucial outcome of the Coaching Method testing was also the detection and elimination of a security issue, allowing for the access to someone else's account without password. The Codiscovery Learning method, which involves users' collaboration, was found suitable for observing the behavior of particular user roles with respect to mutually related accounts. As a result of this collaboration, it was possible to observe the changes related to gaining and removing access to particular accounts, and to verify access restrictions.

The evaluation of system usability conducted through interviews was mainly related to the economical aspects of the system, i.e. to the scope of available ratios as well as the layout and content of the reports generated by the system. As a result of the interviews, some improvements to the reports content have been suggested and introduced. On the other hand, the user feedback, received both through the contact form on the system web page as well as through the tickets mechanism in the system itself was related to the performance of the system regarding the generation of the *.pdf reports. In particular, the feedback revealed some differences in browsers bahavior connected with possible timeouts occuring for very large reports (containing multiple long comments and over 30 charts). As a consequence, reports generation process has been modified and improved to achieve unified and correct behavior across various browsers.

6 Summary

The paper presents a new online service for financial analysis. Its interdisciplinary nature as well as various usage perspectives are described. Selected implementation details are also discussed. In the future the system will be further developed, particularly in terms of its e-Learning capabilities, by the provision of free student's version of the service. Moreover, we plan to extend the scope of financial analysis to also include cash flow statement and industry indicators (as benchmarks).

References

1. Cowen, S.S., Hoffer, J.A.: Usefulness of financial ratios in a single industry. J. Bus. Res. **10**(1), 103–118 (1982)
2. Fabozzi, F.J., Peterson, P.P.: Financial Management and Analysis. Wiley, Hoboken (2003)
3. Pinches, G.E., Eubank, A.A., Mingo, K.A., Caruthers, J.K.: The hierarchical classification of financial ratios. J. Bus. Res. **3**(4), 295–310 (1975)
4. Błach, J., Gorczyńska, M., Wieczorek-Kosmala, M.: Sytuacja finansowa śląskich przedsiębiorstw w dobie kryzysu. CeDeWu. Warszawa (2013)
5. Megginson, W.L., Smart, S.B.: Introduction to Corporate Finance. South-Western, Mason (2006)
6. Alexander, D., Britton, A., Jorissen, A.: International Financial Reporting and Analysis. South-Western, Andover (2011)

7. Baker, H.K., Powell, G.E.: Understanding Financial Management. A Practical Guide. Blackwell Publishing, Oxford (2005)
8. Groppelli, A.A., Nikbakht, E.: Finance. Barron's, New York (2006)
9. Ehrhardt, M.C., Brigham, E.F.: Corporate Finance. A Focused Approach. South-Western, Mason (2011)
10. Baruch, L., Shyam, S.: Methodological issues in the use of financial ratios. J. Account. Econ. **1**(3), 187–210 (1979)
11. McDonald, B., Morris, M.H.: The statistical validity of the ratio method in financial analysis: an empirical examination. J. Bus. Financ. Acc. **11**(1), 89–97 (1984)
12. Deluga, W., Dyczkowska, J.: e-commerce—bezpieczne zakupy. Nierówności Spo3eczne a Wzrost Gospodarczy **22**, 27–38 (2011)
13. Gregor, B., Stawiszyński, M.: e-Commerce. Oficyna Wydawnicza Branta. Bydgoszcz, Poland (2002)
14. Norris, M., West, S.: E-biznes. Wydawnictwa Komunikacji i Łączności. Warsaw, Poland (2001)
15. Coltman, T., Devinney, T.M., Latukefu, A., Midgley, D.F.: e-Business: Revolution, Evolution or Hype? Working Paper 2000/08/MKT. INSEAD Working Paper Series (2000)
16. Beynon-Davies, P.: e-Business maturity and regional development. Int. J. Bus. Sci. Appl. Manage. **2**(1), 9–20 (2007)
17. Kalakota, R., Robinson, M.: e-Business 2.0: Roadmap for Success. Addison-Wesley (2001)
18. Allen, F., McAndrews, J., Strahan, P.: e-Finance: An Introduction. Working Paper 01-36. Wharton Financial Institutions Center (2001)
19. Sejm, R.P.: Ustawa z dnia 29 wrzeonia 1994 r. o rachunkowooci (Accounting Act). http://isap.sejm.gov.pl/DetailsServlet?id=WDU19941210591. Accessed 22 Mar 2016 (1994)
20. Dijkstra, E.W.: Algol 60 translation: An algol 60 translator for the x1 and making a translator for algol 60. Technical Report MR 34/61. Stichting Mathematisch Centrum (1961)
21. Hamblin, C.L.: Translation to and from polish notation. Comput. J. **5**, 210–213 (1962)
22. Saff, D., Cooney, K., Birkner, S., Philipp, M.: JUnit Framework. http://junit.org/junit4. Accessed 28 May 2016 (2006)
23. Ivory, M.Y., Hearst, M.A.: The state of the art in automating usability evaluation of user interfaces. ACM Comput. Surv. **33**(4), 470–516 (2001)

Hierarchical Topic Modeling Based on the Combination of Formal Concept Analysis and Singular Value Decomposition

Miroslav Smatana and Peter Butka

Abstract One of the ways to describe the content of internet sources is known as topic modeling, which tries to uncover the hidden thematic structures in document collections. Topic modeling applied to social networks can be useful for analysis in case of crisis situations, elections, launching a new product on the market etc. It becomes popular research area in recent years and represents the methods to browse, search and summarize large amount of the textual data. The main aim of this paper is to describe a new way for topic modeling based on the usage of Formal Concept Analysis combined with reduction by Singular Value Decomposition of the input data matrix. In difference to other common used method for topic modeling our proposed method is able to generate topic hierarchy, which offer more detail analysis of topics within the collection. Our approach is experimentally tested on the selected dataset of Twitter network contributions.

Keywords Topic modeling · Formal concept analysis · One-sided concept lattices · Singular value decomposition

1 Introduction

In recent years, the resources on the Internet and especially social networks are produced rapidly. Users daily publish the large amount of contributions on social networks, e.g., Twitter (https://twitter.com/) network daily publishes about 340 billions of contributions. These contributions usually reflect user opinions, attitudes on worldwide events, products, persons, etc. One of the areas for automatic analysis of contributions available in digital way is known as topic modeling, which becomes

M. Smatana · P. Butka (✉)
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 04200 Košice, Slovakia
e-mail: peter.butka@tuke.sk

M. Smatana
e-mail: miroslav.smatana@tuke.sk

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_31

more popular with the introduction of dynamic content sources like social networks. It can be very useful not only for browsing the content, but also as an input for other complex tasks with topics which lead to clusters within the collection. The proper understanding and usage of such information can be useful in different types of tasks or analysis, for example:

- Crisis analysis—for example at the time of some war conflict it is possible to monitor how users perceive current situation and if it is necessary to take some appropriate action,
- Launching new product on the market—monitor if users like new product, which bugs it has or which are competitive products according to users,
- Protection of reputation—continuous monitoring of social networks in order to catch contributions with negative opinions on company, organization or person.
- Media—which are the most interesting news, search for reactions of people.

Topic modeling shows us a new way to browse, search and summarize data from different sources using methods which try to uncover hidden thematic structure in data collections. In recent years several methods have been studied and applied for topic modeling from collections of documents. One of the most popular and used method is Latent Dirichlet Allocation (LDA [1]), which was also extended to other variants, e.g., in [2] authors extends LDA to explicitly allow word distributions encoding in order to improve topic alignment quality and synchronization of topics over different languages, Zhai and Graber in [3] proposed the online version of LDA. Also other methods have been proposed like moving average stochastic variation inference [4] or stochastic variational inference [5]. While these methods produce standard topic modeling output, we are more interested in approaches which are able to create not only flat topic classes, but create hierarchical model of particular subtopics. This leads to hierarchical variation of topic modeling, where also some methods were already proposed. We can mention here work of Blei et al. based on nested processes [6], cluster-based abstraction model architecture by Hoffman [7], or hierarchical extension of Dirichlet processes in [8].

Our focus in this paper is to study other approaches which are able to create hierarchical topic models. One of the methods for uncovering the hidden structure within the input textual data is based on decomposition of matrix using Singular Value Decomposition (SVD, see for example [9]). It means that we are able to make reduction of high dimension space and provide representation with factors in lower dimension space. If we select some K best singular values (or K factors), we define more compact input space, i.e., smaller amount of attributes in lower dimension space describe combinations of attributes in high dimension space.

While SVD is good in uncovering the hidden structure of attributes combinations, it is not possible to create hierarchical topics using this method. One of the methods for creation of hierarchy of clusters from input data tables is known as Formal Concept Analysis (FCA [10]). Therefore, we decided to study the possibilities of application of methods from this area to this task. In more details, output of FCA-based methods is hierarchically organized structure of clusters of objects (concepts) according to the shared attributes (based on similarity) of particular objects (within

cluster). Standard approach to FCA is based on 'crisp' case, where object-attribute model (or input formal context) is based on the binary relation (object has/has-not the attribute). Due to fact that most of the data are often described also by other types of attributes, several approaches were designed to work with multi-valued input contexts, e.g., method of conceptual scaling or fuzzy approaches for processing of fuzzy attributes. We can mention work on multi-adjoint lattices [11] or papers by Krajci et al. [12, 13]. In order to provide more practical data mining tool, one-sided fuzzification is used instead of full fuzzy models. It means that only one part of object-attribute model is fuzzified, e.g., objects are considered as it is usual for crisp case (classic subsets) and their attributes obtain fuzzy values (on the side of attributes we have fuzzy sets). Such model is usually called one-sided concept lattice and in [14] its generalized version was introduced (Generalized One-Sided Concept Lattice—GOSCL), which is able to process data tables with different types of attributes, i.e., different truth value structures are used for every attribute.

One of the main problems of FCA-based methods is the large number of produced concepts and therefore large created hierarchy. In order to achieve more comprehensive results, this problem is usually solved by some reduction techniques. In this paper we provide our proposed approach which combines FCA-based method (we used GOSCL extension) with the decomposition of input data table using SVD (as one of the reduction steps) in order to achieve more comprehensive hierarchical topic models. The proposed approach is then compared to standard approaches like LDA and clustering (k-means).

The rest of the paper is organized as follows. In next section we provide preliminaries of methods used within the proposed approach, i.e., some basics related to FCA (GOSCL) and SVD. Next, we describe our proposed approach for combination of these two methods for topic modeling. In the following section some experiments are provided for evaluation of the proposed approach.

2 Preliminaries

In this section we provide basic details regarding the methods which are used in our proposed approach, i.e., details on FCA-based model of Generalized One-Sided Concept Lattices and Singular Value Decomposition (SVD), which are combined in order to find hierarchical topic models from input data.

2.1 Generalized One-Sided Concept Lattices

For the purpose of this paper we only describe some basic details regarding the theory of GOSCL (more details can be found in [14] or [15], some other properties on heterogeneous concept lattices are described in [16, 17]). In general, input data for GOSCL are in the form of object-attribute data table, i.e., 4-tuple (B, A, L, R)

is called *generalized one-sided formal context* with the set of objects B, the set of attributes A, mapping $L : A \rightarrow \mathsf{CL}$ used for identification of truth values structure for particular attributes (CL—class of all complete lattices, for any attribute a is $L(a)$ the structure of values for this attribute). The last element R represents the data table (so-called *generalized incidence relation*), where $R(b, a) \in L(a)$ for any object $b \in B$ and attribute $a \in A$, i.e., $R(b, a)$ is value from structure $L(a)$ assigned to object b for attribute a.

For the creation of Generalized One-Sided Concept Lattice (GOSCL) based on the generalized one-sided formal context (B, A, L, R) we define a pair of mappings $^{\perp} : \mathbf{P}(B) \rightarrow \prod_{a \in A} L(a)$ and $^{\top} : \prod_{a \in A} L(a) \rightarrow \mathbf{P}(B)$ as follows:

$$X^{\perp}(a) = \bigwedge_{b \in X} R(b, a), \tag{1}$$

$$g^{\top} = \{b \in B : \forall a \in A,\ g(a) \leq R(b, a)\}. \tag{2}$$

This pair of mappings $(^{\perp}, ^{\top})$ forms so-called Galois connection between $\mathbf{P}(B)$ and $\prod_{a \in A} L(a)$ and therefore is also basic for forming of resulted concept lattice. Then we can find set $C(B, A, L, R)$ which contains all pairs (X, g), where $X \subseteq B$, $g \in \prod_{a \in A} L(a)$, satisfying $X^{\perp} = g$ and $g^{\top} = X$. Here, X is usually called extent and g intent of some concept (X, g). With the definition of partial order on the set $C(B, A, L, R)$ using formula

$$(X_1, g_1) \leq (X_2, g_2) \quad \text{iff} \quad X_1 \subseteq X_2 \text{ iff } g_1 \geq g_2 \tag{3}$$

we have complete lattice called *generalized one-sided concept lattice*.

Now we provide small illustrative example now. Let $B = b_1, b_2, b_3, b_4, b_5$ are objects and $A = a_1, a_2, a_3$ are attributes. Here, a_1 is binary attribute represented by two element chain with values 0 and 1, attribute a_2 is numeric attribute with values from real interval $< 0, 1 >$ and attribute a_3 is ordinal attribute represented by chain lattice with values from 0, 1, 2, 3. Example of data table is presented in Table 1 and corresponding generalized one-sided concept lattice constructed based on this data table is presented in Fig. 1.

Table 1 Example of input formal context

R	a_1	a_2	a_3
b_1	0	0.0	2
b_2	1	0.5	0
b_3	1	0.0	0
b_4	0	0.4	2
b_5	0	0.2	1

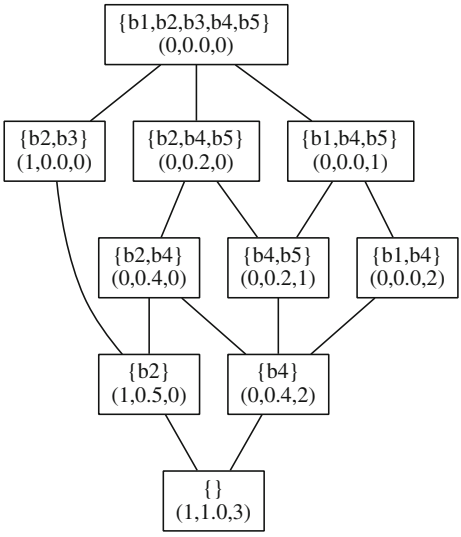


Fig. 1 Example of output GOSCL

Formal concept analysis has been successfully applied in many application. However when it is used for analysis of medium or large collection of data then the number of generated concepts can be large, therefore it is necessary to use some reduction technique for reduction of concepts and/or their connections to make result of analysis more comprehensive. Several reduction methods were already proposed. Some of them are based on the reduction of number of concepts using threshold applied to concepts ranking method—concepts with the rank under the threshold are removed [18, 19]. Interesting paper regarding the possible reductions based on graph properties is described in [20]. Another way proposed in [21] is based on the removing of links (edges) between them in order to create tree-based structure. Reduction of concept lattice can be also achieved by the reduction of input formal context from which concept lattice is generated, example of such reduction with the usage of SVD was presented in [22]. Another possible way is to use clustering methods in order to merge similar concepts into one cluster [23].

2.2 Singular Value Decomposition

Singular value decomposition (SVD [9]) can be characterized from three different points of view. First, it is a way that allows an exact representation of any matrix and easy elimination of less important parts of the representation (factors) to produce an approximate representation. So it can be seen as dimensionality reduction algorithm where number of the reduced dimensions means less accurate approximation of original data. An interesting feature of SVD-based dimensionality reduction is its ability to uncover hidden (thematic) relations between data. It has many useful

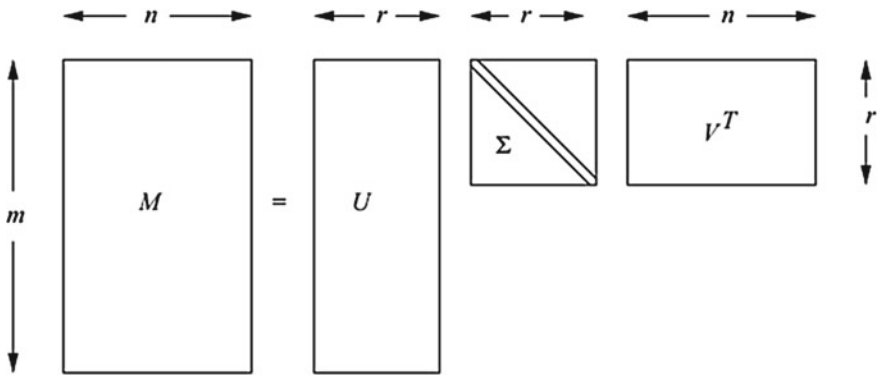


Fig. 2 Singular Value Decomposition of matrix $M = U\Sigma V^T$

applications in signal processing, statistic, machine learning, text mining or information retrieval. Some experiments were also proposed for the combination of standard FCA and SVD in [22].

For simple description of SVD, we can consider matrix M with m rows and n columns and rank r of matrix M. Rank of matrix represents the largest number of linearly independent rows (columns). SVD of M is a factorization of the form $U\Sigma V^T$, where U is an $m \times m$ unitary matrix, Σ is a $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal, and V is an $n \times n$ unitary matrix, the diagonal entries σ_i (representing the factors) of Σ are known as the singular values of M (see Fig. 2). Because the singular values usually quickly decrease we can get only K greatest singular values and make K-reduced SVD of M. One of the possible ways is to directly define K or we can select K based on the energy, defined as sum of the squares of singular values in Σ (e.g., reduce r to K while energy is larger than threshold of original energy, usually it is 90 %).

3 Proposed Approach for Hierarchical Topic Modeling Using FCA and SVD

In this section our proposed approach is described. The basic idea is to combine FCA-based method (GOSCL) with SVD, i.e., SVD is used to uncover hidden thematic structures and FCA is used to create hierarchical topic model from them. The procedure of our proposed approach is shown in Fig. 3.

First step in our procedure is to preprocess input collection of documents (contributions) and the creation of document-term matrix. The preprocessing consists of four steps:

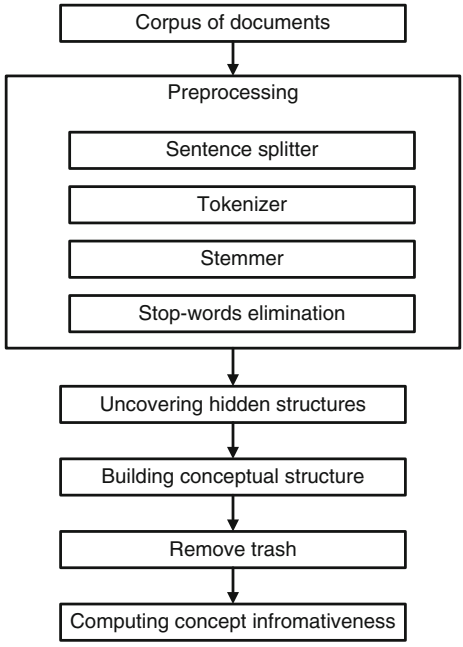


Fig. 3 Singular Value Decomposition of matrix M

- Sentence splitter—separate input documents into sentences,
- Tokenizer—separate input documents into tokens (in our case words),
- Stemming—conversion of input tokens (words) to base form, for our experiments with the corpus of English documents we decided to use Porter stemmer [24],
- Stop-words elimination—remove words with length of 2 or less characters and words which are not meaningful (“and”, “also”, etc.).

Next step is the usage of SVD on input document-term matrix. Due to reduction needs we only use K best singular values. The value of K is selected before next step, which is application of GOSCL algorithm. The result of this step is creation of hierarchy of concepts. The input of GOSCL is matrix U generated by SVD, where rows represent documents and columns represent themes. In order to solve the problem of FCA-based methods and their needs for reduction, we decided to use some simple steps. First, we modify (discretize) the input values of matrix only to values $\{0.0, 0.1, 0.2, \ldots, 1.0\}$. Also, as post-process step of GOSCL we remove the concepts which cover less documents than threshold in percentage of all documents (Remove Trash step). The last step of our approach is the extraction of words informativeness for each concept, which are represented by documents which belongs to the concept and vector of affiliation to topic extracted by SVD. The informativeness is then extracted when concept vector of affiliation is multiplied with the matrix V^T.

Table 2 Comparison of Purity metric achieved by the standard methods (*k*-means and LDA) with our proposed approach (FCA-SVD) for different settings of trash elimination (removing concepts which cover less % of documents than this threshold) and K (best K singular values from SVD reduction step before topic modeling using FCA)

Method	Trash (in %)	K(SVD)	Purity	Number of concepts
LDA	–	–	0.69	4
k-means	–	–	0.76	4
FCA-SVD	5	4	0.73	91
FCA-SVD	5	8	0.69	750
FCA-SVD	5	20	0.55	2942
FCA-SVD	10	4	0.72	68
FCA-SVD	10	8	0.67	455
FCA-SVD	10	20	0.49	784
FCA-SVD	0	4	0.74	185
FCA-SVD	0	8	0.72	1669
FCA-SVD	0	20	0.64	25383

4 Experiments

In this section we describe the experiments with our proposed approach. For this purpose we have used manually sample of manually annotated data with the contributions from Twitter social network, which is free available by the Sanders Analytics.[1] The complete dataset contains 5513 contributions, which discuss four main topics: Apple, Google, Microsoft, Twitter. In our experiments we have used subsample which contains 1000 contributions with the uniform distribution for topics, i.e., every topic was represented by 250 contributions.

The quality of the proposed approach was evaluated according to other standard methods (LDA, *k*-means clustering) and was analyzed from the four points of view:

- Quality of created concepts measured by standard metric—Purity [25],
- Quality of created topic hierarchy—by the graphic visualization of hierarchies using Hasse diagram,
- Number of created concepts,
- Quality of extracted top 15 keywords.

As we can see from Table 2 our approach is able to reach the results of comparable quality with LDA and *k*-means for the Purity metric, when we used lower number of singular values, i.e., the number is near the number of topics (especially if we use 4 and 8 singular values from SVD). There are also other metrics used in standard approach comparisons (like NMI—Normalized Mutual Information), but all of them

[1] http://www.sananalytics.com/lab/twitter-sentiment/.

are not suitable for hierarchical topic models and therefore their comparison is not relevant for our approach.

On the other hand, our approach is able to provide conceptual hierarchical topic model with the comparable quality in purity, what seems to be encouraging result. Of course, as we can see from table, our method is highly dependent from reduction methods and their parameters. If we do not use any reduction method the number of generated concepts will be too high and the hierarchy of concepts will not be suitable for any real application.

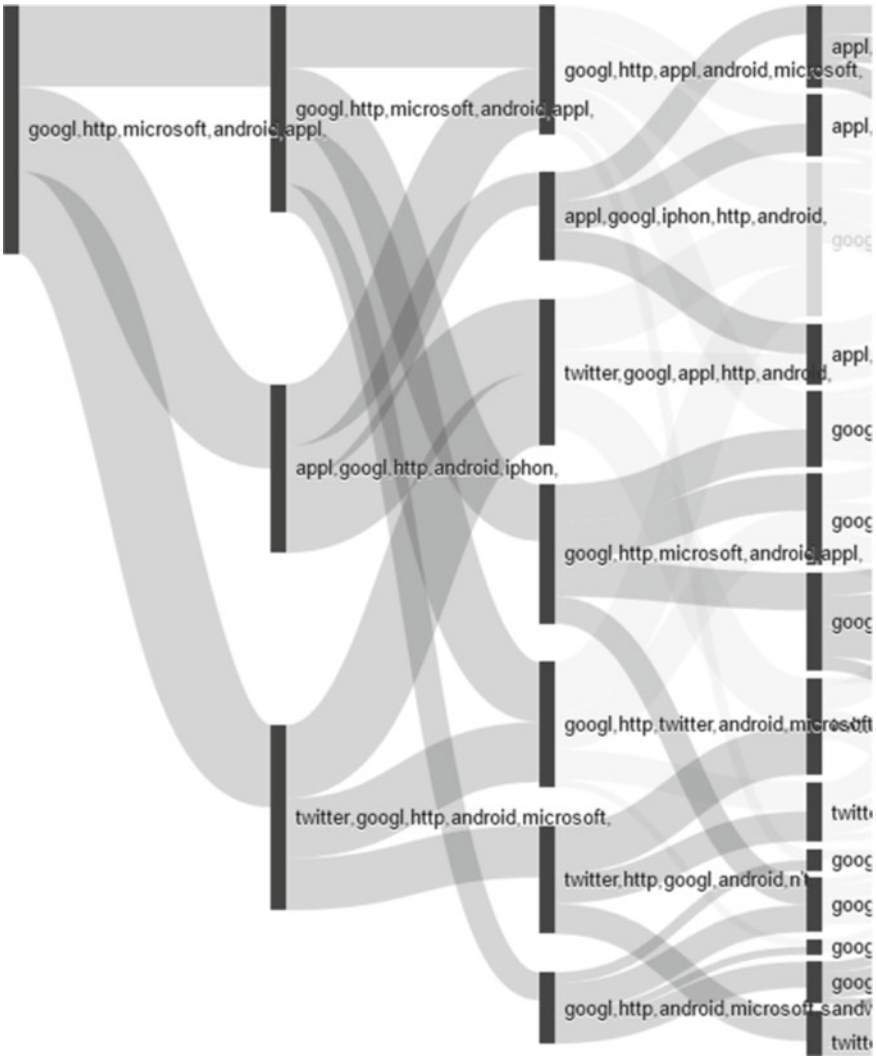


Fig. 4 Visulization of the part of concept hierarchy structure for sample data

appl, iphon, twitter, googl, android, io, love, n't, store, phone, http, great, siri, time, sandwich

googl, android, sandwich, cream, ice, ic, nexus, samsung, http, icecreamsandwich, galaxi, iphon, galaxynexus, app, user

appl, googl, http, android, iphon, twitter, microsoft, sandwich, cream, ice, io, nexus, phone, ic, n't

twitter, http, googl, n't ,follow, facebook, android, day, find, autopilot, hour, second, twittertim, minut, age

googl, http, microsoft, android, appl, twitter, sandwich, cream, ice, nexus, iphon, ic, phone, samsung, n't

Fig. 5 Example of *top* selected words in random concepts from the hierarchy

In Fig. 4 we can see the visualization of the part of generated topic hierarchy with 5 % trash and 4 singular values by SVD. The orientation is from left to right, with the most general concept on left and the more specific concepts on the right. Width of links between nodes (concepts) represents the number of objects in each node. As it is clear from our approach generated concepts contains similar number of objects (contributions). Also our approach is more likely to generate more levels of abstraction with less concepts than less level of abstraction with a lot of concepts. We have also shown the best five terms for concepts in presented visualization. In next Fig. 5 we can see example of best 15 keywords for five randomly selected concepts, from which it is usually possible to identify particular topics and their combinations. Also we would like to adapt our method for processing of data in distributed environment [26, 27], and in different domains like analysis of e-learning user activities [28, 29], but the main focus will be on the experiments with larger social network datasets in order to achieve better evaluation of described method.

5 Conclusions

In this paper we have presented our approach for the creation of hierarchical topic model based on the combination of Formal Concept Analysis (suitable for the creation of topic hierarchy) and Singular Value Decomposition (applied as reduction technique). Our method was tested on selected contributions from Twitter network and showed comparable results with the standard non-hierarchical method in one of the standard metric with the benefit of created topic hierarchy. In the future we would like to extend the experiments on larger datasets, found more comprehensive metric for topic hierarchies and optimize the application of combined methods.

Acknowledgments The work presented in this paper was supported by the Slovak VEGA grant 1/0493/16 and Slovak KEGA grant 025TUKE-4/2015.

References

1. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 694–703 (2003)
2. Petterson, J., Buntine, W., Narayanamurthy, S., Caetano, T., Smola, A.: Word features for latent dirichlet allocation. Adv. Neural. Inform. Process. Syst. **23**, 1921–1929 (2010)
3. Zhai, K., Boyd-Graber, J.: Online latent dirichlet allocation with infine vocabulary. In: Proceedings of ICML 2013, Atlanta, US, pp. 561–569 (2013)
4. Li, X., Ouyang, J., Lu, Y.: Topic modeling for large-scale text data. Front. Electr. Electron. Eng. **16**(6), 457–465 (2015)
5. Hoffman, M., Blei, D., Wang, C., Paisley, D.: Stochastic variational inference. J. Mach. Learn. Res. **14**, 1303–1347 (2013)
6. Blei, D., Griffiths, T., Jordan, M.: The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. J. ACM **57**(2), article number 7, 1–30 (2010)
7. Hofmann, T.: The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In: Proceedings of IJCAI99, Stockholm, Sweden, pp. 682–687 (1999)
8. Paisley, J., Wang, C., Blei, D., Jordan, M.I.: Nested hierarchical dirichlet processes. IEEE Trans. Pattern Anal. Mach. Intell. **37**(2), 256–270 (2015)
9. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, Cambridge (2012)
10. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin (1999)
11. Medina, J., Ojeda-Aciego, M., Ruiz-Calviño, J.: Formal concept analysis via multi-adjoint concept lattices. Fuzzy Set. Syst. **160**, 130–144 (2009)
12. Antoni, L., Krajci, S., Kridlo, O., Macek, B., Piskova, L.: On heterogeneous formal contexts. Fuzzy Set. Syst. **234**, 22–33 (2014)
13. Krajči, S.: A generalized concept lattice. Logic J. IGPL **13**(5), 543–550 (2005)
14. Butka, P., Pócs, J.: Generalization of one-sided concept lattices. Comput. Inf. **32**(2), 355–370 (2013)
15. Butka, P., Pocs, J.: Pocsova: On equivalence of conceptual scaling and generalized one-sided concept lattices. Inf. Sci. **259**, 57–70 (2014)
16. Pocs, J., Pocsova, J.: Basic theorem as representation of heterogeneous concept lattices. Front. Comput. Sci. **9**(4), 636–642 (2015)
17. Pocs, J., Pocsova, J.: Bipolarized extension of heterogeneous concept lattices. Appl. Math. Sci. **8**(125–128), 6359–6365 (2014)
18. Antoni, L., Krajci, S., Kridlo, O.: Randomized Fuzzy Formal Contexts and Relevance of One-Sided Concepts, vol. 9113, pp. 183–199. ICFCA 2015, LNAI (Subseries of LNCS) (2014)
19. Butka, P., Pocs, J., Pocsova, J.: Reduction of concepts from generalized one-sided concept lattice based on subsets quality measure. Adv. Intell. Syst. Comput. **314**, 101–111 (2015)
20. Kardos, F., Pocs, J., Pocsova, J.: On concept reduction based on some graph properties. Knowl. Base Syst. **93**, 67–74 (2016)
21. Melo, C., Le-Grand, B., Aufaure, A.: Browsing large concept lattices through tree ex-traction and reduction methods. Int. J. Intell. Inf. Technol. (IJIIT) **9**(4), 16–34 (2013)
22. Snasel, V., Polovincak, M., Abdulla, H.: Concept lattice reduction by singular value decomposition. In: Proceedings of the SYRCoDIS 2007, Moscow, Russia (2007)
23. Kumar, C.A., Srinivas, S.: Concept lattice reduction using fuzzy k-means clustering. Expert Syst. Appl. **37**(3), 2696–2704 (2010)
24. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)

25. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
26. Sarnovsky, M., Carnoka, N.: Distributed algorithm for text documents clustering based on k-means approach. Adva. Intell. Syst. Comput. **430**, 165–174 (2016)
27. Sarnovsky, M., Ulbrik, Z.: Cloud-based clustering of text documents using the GHSOM algorithm on the GridGain platform. Proc. SACI **2013**, 309–313 (2013)
28. Babic, F., Paralic, J., Bednar, P., Racek, M.: Analytical framework for mirroring and reflection of user activities in e-Learning environment. Adv. Intell. Soft Comput. **80**, 287–296 (2010)
29. Paralic, J., Richter, C., Babic, F., Wagner, J., Racek, M.: Mirroring of knowledge practices based on user-defined patterns. J. Univers. Comput. Sci. **17**(10), 1474–1491 (2011)

RDF Event Stream Processing Based on the Publish-Subscribe Pattern

Dominik Tomaszuk

Abstract In recent years, RDF event streams have been an increasingly widespread data source in a wide range of domains. Existing systems allow us to automatically record pieces of information concerning everyday fast-paced life. This paper proposes methods of representing and processing RDF streams consisting of RDF graphs with time-varying data annotated by an interval. We introduce a graph-based stream model and a solution for RDF streams processing based on the *publish-subscribe* interaction scheme. The initial evaluation on our implementation shows that it has great potential.

1 Introduction and Motivations

In recent years, progress in Web technology has allowed us to automatically record pieces of information of everyday fast-paced life. This process creates vast amounts of online data which grows in an unlimited way. This online data is called a data stream.

Data streams arise in many domains e.g. sensor networks, environmental monitoring, microposts, weblogs, click streams, radio-frequency identification tags, or smart cities. Processing this dynamic information is certainly one of the key challenges for Linked Data.

This paper focuses on representation and processing of streams consisting of RDF elements with time-varying data annotated by an interval. We propose the use of the *publish-subscribe* pattern in RDF stream processing, which allows subscribers to express their interest in an event stream, in order to be notified subsequently of any event stream, generated by a publisher, that matches their registered interest. In our proposal the interacting parties do not need to be participating in the interaction simultaneously and do not need to know each other. This is particularly important

D. Tomaszuk (✉)
Institute of Informatics, University of Bialystok ul, Konstantego
Ciolkowskiego 1M, 15-245 Bialystok, Poland
e-mail: dtomaszuk@ii.uwb.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_32

for federated RDF graph stores supporting streams which appear in thousands of endpoints distributed all over the world.

The paper is constructed according to sections. Section 2 proposes a graph-based stream model, terms connecting to it and a solution for RDF streams processing based on the *publish-subscribe* interaction scheme. Section 3 gives detailed results of our implementation and experiments. Section 4 is devoted to related work. The paper ends with conclusions.

2 A Graph-Based Stream Model

This section defines time in RDF streams, proposes a serialization of RDF streams and shows how it could work in *publish-subscribe* system.

2.1 Time in RDF Streams

In the following subsection we focus on our proposal of RDF streams in the context of named graphs.

The RDF syntax and semantics may be extended to named graphs [1]. The named graph data model is a simple variation of the RDF data model. The basic idea of our proposal is based on a graph naming mechanism.

Definition 1 (*Named graph*) A *named graph* $\mathcal{NG}$ is a pair $\langle u, \mathcal{G} \rangle$, where $u \in \mathcal{I} \cup \mathcal{B}$ is a graph name (Internationalized Resource Identifier or blank node) and $\mathcal{G}$ is an RDF graph.

We define an RDF stream as a sequence of named graphs including associated intervals.

Definition 2 (*RDF stream*) An *RDF stream* S is a linearly ordered sequence of time-annotated named graphs $\langle \mathcal{NG}, \tau \rangle$ where $\mathcal{NG}$ is a named graph and $\tau \in \mathcal{T} \cup \{Nil\}$ is a interval.

The $\mathcal{T}$ time domain is a discrete, ordered infinite set intended to represent discrete time and from now on will be assumed to be the set of natural numbers. The *Nil* time marker indicates a time invariant, used to express a property of the resource that does not change.

Unfortunately, RDF does not support time in statements. There are several proposals [2, 3] that extended RDF but current implementations do not support this. We propose time vocabulary (TV) to be compatible with existing implementations. Below we present the RDF schema of TV.

```
tv:end rdf:type rdf:Property ;
       owl:equivalentProperty tl:end ;
       rdfs:domain :Interval ;
       rdfs:range xsd:dateTimeStamp .
tv:start rdf:type rdf:Property ;
         owl:equivalentProperty tl:start ;
         rdfs:domain :Interval ;
         rdfs:range xsd:dateTimeStamp .
```

Our properties can be mapped to the Timeline ontology.[1]

2.2 Serialization

In the following subsection we focus on our proposal of RDF serialization and punctuation in the stream.

The abstract model can be represented in different serializations, which support RDF streams including our proposal (see Definition 2). We propose a new SJSON-LD serialization for RDF streams based on JSON-LD [4].

Unfortunately, the same RDF graphs and named graphs in JSON-LD can exist in different forms. To solve this problem we propose Algorithm 1. The algorithm converts RDF data to JSON-LD without context key, which is in a uniform structure.

input : RDF data D
output: exp-JSON-LD
1 create Promise P; // see [5]
2 $J \leftarrow$ enrichWithIntervals(toJSONLD(D));
3 $P \leftarrow$ flatten(expand(J)); // see [6]
4 **return** P;

Algorithm 1: SJSON-LD generation

A receiver of an RDF stream should know that all statements belonging to the RDF graph are transmitted in order to start processing it. We propose using Record Separator ASCII control code to punctuate pieces of data.

Definition 3 (*Punctuation*) A *punctuation n* is a pattern inserted into the RDF stream with the meaning that no data streams from named graph will occur further on in the stream.

Definition 4 (*SJSON-LD*) Assume that value v is `true`, `false`, `null`, number from the set of real numbers $\mathbb{R}$, string from the set of literals $\mathcal{L}$, array or key/value pair, a is the array $\langle v_1, v_2, \ldots v_n \rangle$ and p is the key/value pair $\langle k, v \rangle$, where $k \in \mathcal{L}$ is called the key. A *SJSON-LD* j is defined as $j = \langle a_1, n, a_2, n, \ldots, a_i, n \rangle$, where $i \geq 1$.

[1] http://purl.org/NET/c4dm/timeline.owl#.

Below we present the grammar of SJSON-LD in EBNF.[2]

```
SJSON-LD     ::= (exp-JSON-LD RS)*
RS           ::= %x1E
exp-JSON-LD ::= <see Algorithm 1>
```

There is no need to use TV in our SJSON-LD because we propose `@start` and `@end` keys to associated intervals. These JSON keys should be included in the same level as `@id`.

We also propose a method for transformation from our approach to the named graph model. It is important because of compatibility with existing systems. Algorithm 2 presents the process of transformation from SJSON-LD, which uses named graphs.

input : array of SJSON-LD A
output: $\mathcal{NG}$
1 $c \leftarrow$ getContext(A);
2 $ng \leftarrow$ createNamedGraph(c);
3 $sdts \leftarrow$ getTime(A,"`@start`");
4 $edts \leftarrow$ getTime(A,"`@end`");
5 setTriple(ng, c, "`tv:start`", $sdts$`^^xsd:dateTimeStamp`);
6 setTriple(ng, c, "`tv:end`", $edts$`^^xsd:dateTimeStamp`);
7 **foreach** $a \in A$ **do**
8 (s, p, o) $\leftarrow$ getTriple(A);
9 setTriple(ng, s, p, o);
10 **return** ng;

Algorithm 2: Transformation from SJSON-LD to named graphs

2.3 *Publish-Subscribe System*

In the following subsection we focus on our proposal of the *publish-subscribe* system for RDF event stream processing.

We propose the use of the *publish-subscribe* pattern in RDF stream processing, which allows subscribers to express their interest in an event stream, in order to be notified subsequently of any event stream, generated by a publisher, which matches their registered interest.

Definition 5 (*RDF publishers and subscribers*) *RDF publishers* (so-called producers) publish RDF data on an *event stream manager*[3] and RDF subscribers (so-called consumers) subscribe to the data that they want to receive from that manager.

[2] http://www.w3.org/TR/REC-xml/#sec-notation.

[3] In *publish-subscribe* pattern sometimes it called a message broker.

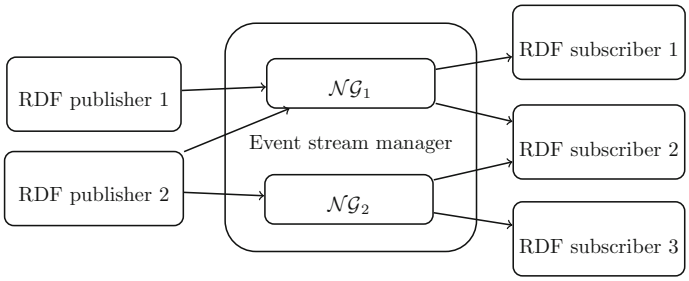


Fig. 1 A simple graph context-based publish/subscribe proposal

Definition 6 (*Notification*) The act of delivering an RDF event stream is denoted by *notification*.

We notice that RDF subscribers are usually interested in particular RDF event streams, and not in all event streams. Our proposal uses named graphs as a subscription scheme in the *publish-subscribe* pattern. We increase the notion of channels with methods to classify graph stream context. We present our proposal in Fig. 1. An RDF publisher calls a `publish()` operation, which generates an RDF event. Then, the event stream manager sends the RDF event to all relevant subscribers. On the other hand, an RDF subscriber registers her/his interest in RDF events using graph context and she/he calls a `subscribe()` operation on the event stream manager. Note that subscribers do not know the effective sources of these events.

Definition 7 (*Publish function*) A *function publish* $: \mathcal{I} \rightarrow \mathcal{NG}$ is the mapping from an IRI (Internationalized Resource Identifier) to a named graph. It should provide data in SJSON-LD format.

Definition 8 (*Subscribe function*) A *function subscribe* $: \mathcal{NG} \rightarrow \mathcal{I}$ is the mapping from a named graph to an IRI. It should provide data in SJSON-LD format.

Subscribers can cancel subscription by calling `unsubscribe()` operation, which is opposite to `subscribe()` operation.

Definition 9 (*Stream sources and destinations*) A *stream source* $S_s \subset \mathcal{I}$ is the set $S_s = \{s \in \mathcal{I} : publish(s) \neq \emptyset\}$. A *stream destination* $S_d \subset \mathcal{I}$ is the set $S_d = \{d \in \mathcal{I} : subscribe(d) \neq \emptyset\}$.

This approach guarantees unlimited support for time, space and synchronization decoupling. It means that the interacting parties do not need to be participating in the interaction simultaneously and do not need to know each other. Moreover, publishers are not blocked while producing RDF event streams and subscribers can get asynchronously notified of the occurrence of an RDF event stream.

3 Implementation and Evaluation

In the following section we focus on implementation experience of RDF event stream *publish-subscribe* system. We also present performance experiments.

Our testbed uses the following components: Virtuoso[4] quadstore, SJSON-LD encoder connected to Virtuoso via ODBS, event stream manager with a named graph broker, and client based on Jena[5] and SJSON-LD parser.

SJSON-LD encoder, SJSON-LD parser and event stream manager are written in Python. A client, which plays a role of RDF subscriber, is implemented in Java. A named graph broker uses MongoDB[6] to store data about subscription and other metadata. An event stream manager supports `publish()`, `subscribe()` and `unsubscribe()` operations.

We define function calls that capture the core of our system:

- `publish(rdf_stream_event)`
- `handle` ← `subscribe(named_graph_filter, expires)`
- `unsubscribe(handle)`

The experimental environment was run on a Linux Mint 17.1 LTS. Tests were performed on computers with an Intel Core i7-4770K CPU @ 3.50GHz and 8 GB of RAM. The network traffic was limited only to communication between testbed components.

To test our environment we prepared 12 datasets about bike-sharing. We established that update frequency is [65000; 780000] statements per minute, so that the first iteration has 5000 events, and the last iteration has 60000 events.

In the first step, we measured the performance of encoding SJSON-LD on the RDF producer side. The first plot in Fig. 2 presents the results of the experiment, in which we generate RDF events. The plot shows the arithmetic mean encoding time from 10 runs. We analyze three cases: local encoding, encoding and sending to event stream manager via HTTP, and encoding and sending to event stream manager via HTTPS. The plot shows that times are nearly quadratic to the number of RDF events. Coefficients of determination are $R^2 \approx 0.996$ in local encoding, $R^2 \approx 0.998$ in communication via HTTP protocol and $R^2 \approx 0.997$ in communication via HTTPS protocol. The results are not surprising; the local generating is the fastest one and encoding with sending via HTTPS is the slowest one.

Then, we measured the performance of parsing SJSON-LD on the RDF subscriber side. The second plot in Fig. 2 presents the results of the experiment, in which we process RDF events. The plot shows the arithmetic mean encoding time from 10 runs. We analyze three cases: local parsing, receiving from and parsing, and receiving from HTTPS and parsing. The plot shows that times are nearly logarithmic to the number of RDF events in local parsing. Times are nearly linear to the number of RDF events in other cases. Coefficients of determination are $R^2 \approx 0.936$ in local parsig,

[4] http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/.

[5] https://jena.apache.org/.

[6] http://mongodb.org/.

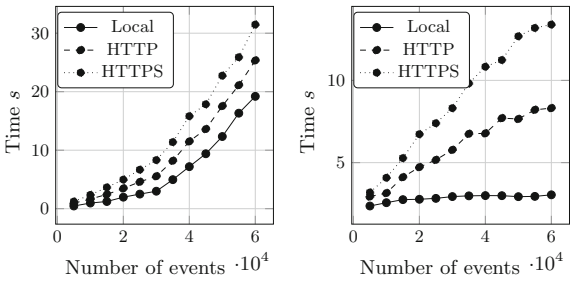


Fig. 2 Encoding and parsing result of performance measurement based on the number of events

$R^2 \approx 0.975$ in communication via HTTP protocol and $R^2 \approx 0.988$ in communication via HTTPS protocol. Just as in the previous case the local parsing is the fastest one and parsing with secure communication is the slowest one.

4 Related Work

In the context of the Semantic Web there are new proposals for publishing data streams. The paper [7] proposes an approach to publish data streams as Linked Data. In [8] the concept of Linked Stream Data is defined, a way in which the Linked Data principles can be applied to publishing data streams. In [9] an ontology-based approach for providing data access and query capabilities to streaming data sources are proposed. More general approaches are presented in [10, 11]. In [10] authors provide an analysis of the experimental outcomes for improving current engines. In [11] a continuous query processor and a reasoner that can deal with streaming data are proposed. The paper [12] surveys existing approaches and proposals in the area of data streams.

In the Semantic Web there are also a number of publications related to RDF stream processors and query languages, e.g. CQELS [13], EP-SPARQL [14, 15], C-SPARQL [16], SPARQL$_{stream}$ [17], TA-SPARQL [18], Sparkwave [19], INSTANS [20], and Streaming SPARQL [21]. The paper [20] concentrates on event processing by supporting events in datasets. Papers [13, 16, 17, 19] describe data stream processing with special extensions for window extraction. The papers [14, 15] propose time interval operators and sequence. In this area there are also papers that focus on reasoning on streams, e.g. [22, 23].

On the other hand, in distributed applications there are several categories of messaging patterns, like a request-reply, or a Representational State Transfer. A request-reply is pattern that connects a set of clients to a set of services. An example of this pattern is Common Object Request Broker Architecture [24], which gives an abstraction for building complex distributed communication. Another example is Remote Procedure Call [25], which is type of communication between two points over a net-

work. XML-RPC[7] and JSON-RPC[8] are variants of Remote Procedure Call, which use concrete syntax. Simple Object Access Protocol [26] evolved as a successor of XML-RPC, though it borrows its transport. The last pattern is a Representational State Transfer [27], which is very specific because of the lack of a specified syntax. This approach is a coordinated set of constraints applied to the design of components in a distributed system.

There are also P2P-based Publish/Subscribe approaches for RDF data storage [28–31]. In [28] Cai et al. propose a content-based Publish/Subscribe layer. Chirita et al. [29] show a set of algorithms to manage publications and subscriptions on top of super-peer. In [30] Liarou et al. present a content-based Publish/Subscribe system, which uses conjunctive multi-predicate queries. Ranger et al. [31] propose an information sharing system that is a topic-based Publish/Subscribe platform. Unfortunately, none of these proposals support RDF streams.

In the context of the World Wide Web there are also web feed solutions, which focus on providing users with frequently updated content. Good examples are RSS[9] and Atom [32]. The idea of RSS is to publish frequently updated blog entries, news headlines, multimedia, etc. Some versions of RSS support RDF syntax. Atom evolved as a successor of RSS, but do not support RDF syntax.

5 Conclusions and Further Work

The topic of RDF data streams is a very recent one. Data streams are an increasingly widespread data source in a wide range of domains. Moreover, RDF streams from different domains could mean quite various things. We argue that RDF event streams nowadays face some of the challenges we were facing five years ago, when Linked Data was at its infancy.

In this paper, we provided a proposal for RDF event stream processing and proposed a graph-based stream model. In the next step, we introduced a new serialization called SJSON-LD based on the model, and algorithms to generate and transform to SJSON-LD. Our approach is based on SJSON-LD and the *publish-subscribe* pattern, which allows subscribers to express their interest in RDF event streams. We have proposed named graph filtering as the best solution for specifying the events in Linked Data environment. The advantage of our approach is unlimited support for time, space and synchronization decoupling. The initial evaluation on our implementation shows that it has great potential.

Future work will focus on optimizing routing in the event stream manager. To solve this problem we will consider introducing an `advertise()` operation, which announces intention of publishing certain RDF event streams.

[7] http://xmlrpc.scripting.com/spec.html.

[8] http://www.jsonrpc.org/specification.

[9] http://web.resource.org/rss/1.0/spec.

References

1. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In: WWW '05: Proceedings of the 14th International Conference on World Wide Web, pp. 613–622. ACM (2005)
2. Gutierrez, C., Hurtado, C.A., Vaisman, A.: Temporal rdf. The Semantic Web: Research and Applications, pp. 93–107. Springer (2005)
3. Udrea, O., Recupero, D.R., Subrahmanian, V.S.: Annotated rdf. ACM Trans. Comput. Logic (TOCL) **11**(2), 1–41 January (2010)
4. Lanthaler, M., Sporny, M., Kellogg, G.: JSON-LD 1.0. W3C recommendation, Wide World Web, January (2014)
5. Kambona, K., Boix, E.G., De Meuter, W.: An evaluation of reactive programming and promises for structuring collaborative web applications. In: Proceedings of the 7th Workshop on Dynamic Languages and Applications, pp. 3:1–3:9. ACM (2013)
6. Lanthaler, M., Kellogg, G., Sporny, M.: JSON-LD 1.0 Processing Algorithms and API. W3C recommendation, Wide World Web, January (2014)
7. Barbieri, D.F., Valle, E.D.: A proposal for publishing data streams as linked data. In: Linked Data on the Web Workshop (2010)
8. Sequeda, J.F., Corcho, O.: Linked stream data: A position paper (2009)
9. Calbimonte, J.-P., Jeung, H.Y., Corcho, O., Aberer, K.: Enabling query technologies for the semantic sensor web. Int. J. Semant. Web Inf Syst, **8**(EPFL-ARTICLE-183971), 43–63 (2012)
10. Le-Phuoc, D., Dao-Tran, M., Pham, M.-D., Boncz, P., Eiter, T., Fink, M.: Linked stream data processing engines: Facts and figures. The Semantic Web–ISWC 2012, pp. 300–312. Springer (2012)
11. Walavalkar, O., Joshi, A., Finin, T., Yesha, Y.: Streaming knowledge bases. In: International Workshop on Scalable Semantic Web Knowledge Base Systems (2008)
12. Margara, Alessandro, Urbani, Jacopo, van Harmelen, Frank, Bal, Henri: Streaming the web: reasoning over dynamic data. Web Seman. Sci. Serv. Agents World Wide Web **25**, 24–44 (2014)
13. Le-Phuoc, D., Dao-Tran, M., Parreira, J.X., Hauswirth, M.: A native and adaptive approach for unified processing of linked streams and linked data. The Semantic Web–ISWC 2011, pp. 370–388. Springer (2011)
14. Anicic, D., Fodor, P., Rudolph, S., Stojanovic, N.: EP-SPARQL: a unified language for event processing and stream reasoning. In: Proceedings of the 20th International Conference on World Wide Web, pp. 635–644. ACM (2011)
15. Anicic, Darko, Rudolph, Sebastian, Fodor, Paul, Stojanovic, Nenad: Stream reasoning and complex event processing in ETALIS. Semantic Web **3**(4), 397–407 (2012)
16. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Querying rdf streams with c-sparql. ACM SIGMOD Rec. **39**(1), 20–26 (2010)
17. Calbimonte, J.-P., Corcho, O., Gray, A.J.G.: Enabling ontology-based access to streaming data sources. The Semantic Web–ISWC 2010, pp. 96–111. Springer (2010)
18. Rodrıguez, A., McGrath, R. Liu, Y., Myers, J., Urbana-Champaign, I.L.: Semantic management of streaming data. Proc. Seman. Sensor Netw. **80**, (2009)
19. Komazec, S., Cerri, D., Fensel, D.: Sparkwave: continuous schema-enhanced pattern matching over rdf data streams. In: Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems, pp. 58–68. ACM (2012)
20. Rinne, M., Nuutila, E.: Constructing event processing systems of layered and heterogeneous events with SPARQL. On the Move to Meaningful Internet Systems: OTM 2014 Conferences, pp. 682–699. Springer (2014)
21. Bolles, A., Grawunder, M., Jacobi, J.: Streaming SPARQL-extending SPARQL to Process Data Streams. Springer (2008)
22. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Incremental Reasoning on Streams and Rich Background Knowledge. Springer (2010)

23. Barbieri, D., Braga, D., Ceri, S., Della Valle, E., Huang, Y., Tresp, V., Rettinger, A., Wermser, H.: Deductive and Inductive Stream Reasoning for Semantic Social Media Analytics (2010)
24. Siegel, J.: CORBA 3 Fundamentals and Programming, vol. 2. Wiley (2000)
25. Ananda, A.L., Tay, B.H, Koh, E.-K.: A survey of asynchronous remote procedure calls. ACM SIGOPS Operating Syst. Rev. **26**(2), 92–109 (1992)
26. Moreau, J.-J., Mendelsohn, N., Gudgin, M., Hadley, M., Nielsen, H.F., Lafon, Y., Karmarkar A.: SOAP Version 1.2 Part 1: Messaging Framework (Second Edition). W3C recommendation, Wide World Web, April (2007)
27. Fielding, R.T., Taylor, R.N.: Principled design of the modern Web architecture. ACM Trans. Internet Technol. (TOIT) **2**(2), 115–150 (2002)
28. Cai, Min, Frank, Martin R., Yan, Baoshi, MacGregor, Robert M.: A subscribable peer-to-peer RDF repository for distributed metadata management. J. Web Sem. **2**(2), 109–130 (2004)
29. Chirita, P.-A., Idreos, S., Koubarakis, M., Nejdl, W.: Designing Semantic Publish/Subscribe Networks Using Super-Peers (2006)
30. Liarou, E., Idreos, S., Koubarakis, M.: Publish/subscribe with RDF data over large structured overlay networks. Databases, Information Systems, and Peer-to-Peer Computing, pp. 135–146. Springer (2007)
31. Ranger, D., Cloutier, J.-F.: Scalable peer-to-peer rdf query algorithm. In: WISE Workshops, Lecture Notes in Computer Science, vol. 3807, pp. 266–274. Springer (2005)
32. Nottingham, M., Sayre, R.: The Atom Syndication Format. RFC 4287, Internet Society, December (2005)

Influence of Message-Oriented Middleware on Performance of Network Management System: A Modelling Study

Krzysztof Grochla, Mateusz Nowak, Piotr Pecka and Sławomir Nowak

Abstract Gathering data from Internet of Things and management of IoT devices requires an efficient communication architecture. In this paper we analyse architectures of the scalable, sensor-oriented IoT network management system, as well as the pros and cons of introducing into it a message-oriented middleware server (message broker). We compare two architectures: with distributed buffers and with a centralized message broker. The analysis was conducted on the basis of Markov chains and discrete event simulation.

Keywords IoT management · Network modeling · Markov chains · Discrete event simulation

1 Introduction

Networks of smart sensors or meters, passing data from physical objects to the digital world—from climate and weather parameters in a given area to the measurement of water and gas usage—are more widely used. The difficulty lies in creating an effective management system able to receive and process messages from thousands or millions of devices.

The paper aims at establishing the pros and cons of using a message-oriented middleware (MOM) messaging broker in the management system of sensor-centric

K. Grochla · M. Nowak (✉) · P. Pecka · S. Nowak (✉)
Institute of Theoretical and Applied Computer Science,
Polish Academy of Sciences, Gliwice, Poland
e-mail: mateusz@iitis.pl

K. Grochla
e-mail: kgrochla@iitis.pl

P. Pecka
e-mail: piotr@iitis.pl

S. Nowak
e-mail: emanuel@iitis.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_33

IoT network, handling millions of messages coming from a cloud of smart devices. The system can be viewed as an extension of the model presented in [1] towards high scalability of IoT network management system. The results of this work have been used by Proximetry to determine design choices in the practical implementation of a large IoT system [2]. Proximetry is a company which delivers software solutions to manage the most critical aspects in the IoT, from millions of remote sensing devices.

The management system is responsible for handling numerous tasks. Particular tasks, resulting from messages coming from the network or operator actions are of different types and require different handling. To make the scaling of a network possible, each task is handled by a different *functional server*. For modelling purposes each functional server appears in the system in one instance only, however in the real system each of them could be replicated. Proposed system architecture is very flexible in terms of scalability thanks to gateways (GW) executing routing and load-balancing tasks. One can also consider adding MOM as a messaging broker (MB)—and intermediate entity aimed at queuing and routing messages among the devices and servers. The introduction of MB into the management system has also been justified in [1].

MOM can be implemented by means of AMQP (Advanced Message Queuing Protocol) [3]. AMQP is an open-standard protocol that allows the network applications to connect with each other. AMQP supports queueing and routing of messages, including point-to-point and pub-sub messaging. There are also other message systems, eg. MSMQ (Microsoft Message Queuing) [4] or JMS (Java Message Service) [5], but only AMQP is an open standard, wire-level protocol, defining format of the data that is sent over the network. Therefore its use is not limited to any particular software vendor or programming language. Its features, along with efficient and well-established open-source implementations like RabbitMQ and ApacheMQ [6], make AMQP a widely adopted choice as messaging middleware. The RabbitMQ makes it possible to assure high availability and has proven to provide high performance [7].

In this paper we present the distributed architecture of the management system of big (several thousands of nodes) network of IoT sensors, and the results of comparing two variants of its design—with and without a messaging broker. The study was conducted by simulation, and the correctness of the model was proven by means of a Continuous Time Markov Chains modelling analysis. AMQP was used as a messaging protocol, but the results can be applied also when another MOM system is to be used.

The use of AMQP based MOM was not widely presented in research. In [8] authors presented a modelling study of an event notification and intelligent inference system, using an AMQP and open-source messaging broker, achieving a high degree of parallelism and scalability. [9] presents model-checking and time logic models of message brokers in pub-sub networks. Psiuk et al. ([10]) performed real-world (not modelling) tests of architecture with MOM, however they also made no comparison of network with and without it. Therefore our contribution, which focuses on the influence of an AMQP message broker on the network performance, is unique.

1.1 Elements of the System

The proposed system contains functional units—dedicated network devices extracted during the conceptual work as elements necessary to achieve the required functionality. The architecture shown below presents a typical three-layer cloud system implementing the publish-subscribe design pattern. The list of servers includes the following units (parentheses contain abbreviations used later in the text):

- Load Balancer (LB) leads the network traffic coming from smart devices to the GW, so as to maintain a balanced load on GW units and thus their maximum performance.
- Gateway (GW), which is an intermediate node, multiplexes packets from smart devices and leads them to the appropriate servers, and also in the opposite direction—heading the notifications from servers to the corresponding devices in a network. There are more GW devices in the system due to the expectation of an excessive traffic.
- Functional servers (FS1 ... FSn), performing tasks like billing, statistics data generation, handling of logs and alarm messages and other tasks specific for particular network
- Visualization Server (VS), which prepares overview and errors data for visualisation

Messages from smart devices come via LB to GW, where they are routed to particular servers FS1..FSn. Usage of load balancer is not common in IoT systems, but it frees network designer from having to manually assign nodes to the gateways, and versatility of the system was one of its key requirements. LB does not analyse content of incoming messages, so it is able to process messages fast enough. The servers analyse the state of the network and pass synthetic information to the VS device, sharing this information with the human operator. The operator can influence the status of the network also through VS, passing appropriate commands, translated into network messages and routed to the appropriate servers. The server receiving operator command, in turn, generates a frame or a series of frames, which shall be addressed to the GW. GW forwards the frame to the corresponding devices on a network. Because the network management system is prepared to handle a large number of messages of different types and with variable intensity, it was necessary to introduce queues of messages waiting to be processed by the units. Two versions of the system architecture were tested, managing queues messages flowing between individual GWs and servers in different way.

1.2 Architecture Based on Direct Message Passing

The first version of the system architecture, shown on Fig. 1, assumed that message queues are handled by individual servers. Messages coming from the device are sent

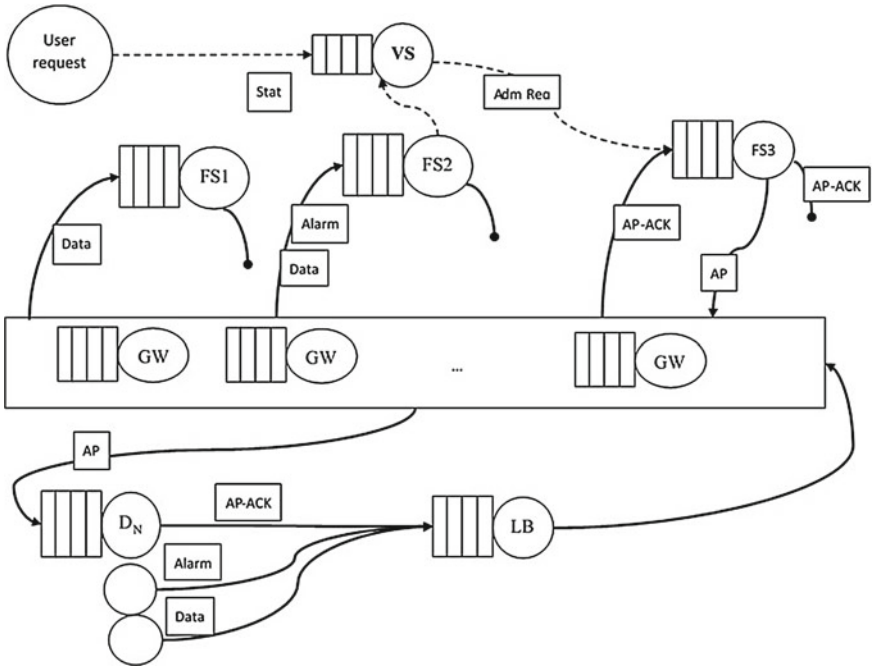


Fig. 1 Basic queueing model of smart devices network management system

to the LB queue separating them into individual GW queues. GW gets another frame from its queue, processes and sends them directly to the appropriate server queue. Also, messages sent by the servers go to the appropriate queue—both in VS as well as GW devices.

1.3 Architecture Based on Centralized Message Broker

In the second version of the architecture (Fig. 2) a specialized unit for queue management—message broker—is introduced. Its task is storage (caching) messages flowing between functional servers, VS and GWs. The purpose of using MOM was to relieve the individual servers from message queues management tasks, allowing more resources (memory and CPU time) to be available for the basic tasks performed by individual servers. On the other hand, the path of network message is longer, as it must now go through an additional device. In the model MB was implemented as two entities—one (FS-MB) for queueing messages from GWs to FSs, another (GW-MB)—for the opposite direction.

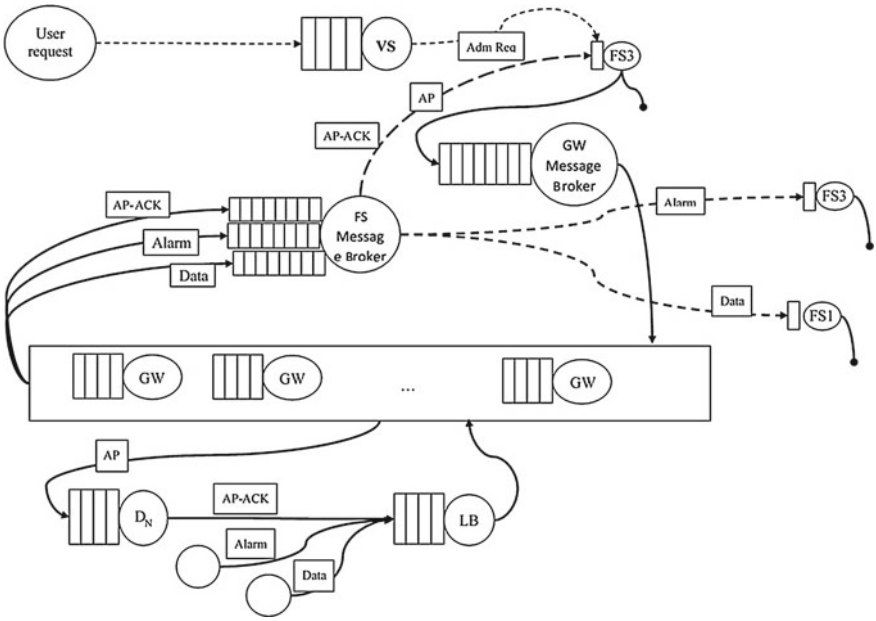


Fig. 2 Network management system with Message Broker

2 Modelling

And objective of the research was to compare two architectures running the same software and hardware elements, based on some realistic assumptions. The parameters of the model are based on authors' experience with implementation of similar systems (in Java, C#and Erlang), general description of the system and performance metrics of RabbitMQ architecture. The requirements and experience of Proximetry engineers have also been taken into account.

- Network of devices consisting of 10 mln devices (in this case, the energy meters, connected by the ad-hoc network), each sending a data message every 900 s, was modelled as device sending a message with exponential distribution of sending periods and average time between messages $t = 90\mu s$
- Load balancer was modelled as a device having constant service time of a packet $t = 16.67\mu s$
- GW service time, based on real implementation of similar server in Java, was modelled as follows:
  - Probability $p = 66.666666\,\%$, that service time has constant distribution with $t = 50\mu s$
  - Probability $p = 33.333333\,\%$ that service time has exponential distribution with average $t = 500\mu s$

  - Probability $p = 0.000001\,\%$ that service time has linear distribution from range 1..100000ms, comes from periods of work of Java garbage collector.

- FS1 functional server is able to handle 26 000 messages/s, what was modelled as exponential service time with average $t = 38\mu s$
- FS2 functional server was modelled as server with service time having exponential distribution with average time $10ms$
- Message Broker works in two modes from the simulation point of view, so was modelled as two independent servers. In communicating with functional servers the sending speed is strictly adjusted to server speed, so the control mechanism of NextMessage frames was introduced in the model to reflect it. In communication with GW, which maintain a classical model of their buffers, service time having exponential distribution was used (average time $t = 10\mu s$)

To examine the influence of a messaging broker into the system, a series of simulation studies has been performed. Simulation can reflect the work of the system in the most detailed way, but, if the physical system is not built, there is no way to directly verify the correctness of the simulation model. Therefore models of the same system, built independently for other modelling methods, are used for the indirect verification of simulation models.

We used the Continuous Time Markov Chains (CTMC) [11] method to verify the correctness of the model. CTMC belongs to analytical modelling methods and provides modelling results by calculating mathematical formulas rather than by simulation.

2.1 Markovian and Discrete Event Simulation Models

The markovian analysis of the queuing model for the evaluation of performance was conducted in the tool Olymp-2 [12], created in IITiS PAN for testing performance computer systems and networks with CTMC. Because the size of the test system is very large in terms of technical capability analysis by Markov chains, the system is divided into fragments, and the size of queues is reduced in comparison to the actual implementation.

The results obtained were verified using discrete event-driven simulator OMNeT++. The satisfactory accuracy of the results of the markovian modelling and simulation modelling made it possible to assume that the results are reliable, and that the models (created independently for both methods) were built in the correct way. This in turn made it possible to proceed to the next step of the simulation modelling. As discrete event simulators do not have such limitations in the size of the model as the markovian models, we prepared a complete model management system, in which the verified model fragments were combined, and the queues extended to the actual size. Since there were no changes to the code that describes the operation of the different modules, but only the parameters and connections between modules were changed, it could be assumed that the reliability of the simulation remains the same.

3 Results

Due to the limitations of the Markov software, coming from "explosion of states" phenomena typical of MC modelling, Markovian modelling was run for the limited size of the model, and the same parameters were applied to the simulation model. Obtaining the same results confirmed the correctness of the simulation model and made it possible to change its parameters—mainly queue lengths—to realistic values. Results obtained allowed us to draw conclusions about the influence of implementing MOM within management system described in the paper on its performance.

3.1 Comparison of Markov and Simulation Results

The aim of markovian modelling was to prove the correctness and reliability of the simulation model. The Tables 1, 2, 3 and 4 present the distribution of probability of queue length for particular nodes in the model, both in a system with and without a Message Broker. The parameters of the model—queue lengths and intensity of messages coming from the network are lower than in the full simulation model, so the results are not useful for judging the system in any way.

Exemplary results for three Markov scenarios are presented, to show the accuracy of matching of both models. In test scenarios two types of messages are transmitted—data messages and alarms. Arrival rate for data $\lambda_{Data} = 1$, for alarms $\lambda_{Alarm} = 0.5$. Service rate of GW, MB nad FS2 are respectively $\mu_{GW} = 1$, $\mu_{MB} = 0.5$, $\mu_{FS2} = 1$. Two GWs are present in the system. All messages queues are of length 3. In **Markov Scenario 1** Alarm and Data messages in system with no Message Broker (Data messages to FS1 and Alarms to FS2). In **Markov Scenario 2** Alarm only messages in system with no Message Broker. In **Markov Scenario 3** Alarm only messages in system with Message Broker.

The tables and charts show the probability of a particular client number in the queue. Due to the limited scope of this paper only some of the results are presented, however all obtained results show the same level of similarity between simulation and Markov modelling.

Table 1 Distribution of queue lengths for node GW1 in scenario 1

Clients	Probability	
	Markov	*Simulation*
0	0.85182296	0.851823
1	0.127680877	0.127681
2	0.018434011	0.018434
3	0.002062152	0.002062

Table 2 Distribution of queue lengths for node FS2 in scenario 1

Clients	Probability	
	Markov	*Simulation*
0	0.538022794	0.538022805
1	0.260745586	0.260745553
2	0.127862231	0.127862281
3	0.073369389	0.073369361

Table 3 Distribution of queue lengths for node GW1 in scenario 2

Clients	Probability	
	Markov	*Simulation*
0	0.69834866	0.698349
1	0.209165244	0.209165
2	0.068185134	0.068185
3	0.024300962	0.024301

Table 4 Distribution of queue lengths for node GW1 in scenario 3

Clients	Probability	
	Markov	*Simulation*
0	0.500344282	0.500344
1	0.283112102	0.283112
2	0.138091065	0.138091
3	0.078452551	0.078453

3.2 Simulation Results

An exemplary scenario typical of network management was selected, as it includes sending order messages to particular managed devices, and receiving answers from them. It comprises the Administration Packet (AP) sent from the functional server FS3 to: (Scenario 1) 2 000 000 devices out of 10 000 000 in the network, (Scenario 2) 5 000 000 devices out of 10 000 000 in the network, (Scenario 3) 10 000 000 devices out of 10 000 000 in the network, and its acknowledgement packet back to FS3. In every scenario Data messages, carrying basic data from a device to the FS1 server are sent by all smart devices on a synchronous basis—every device sends a message every 900 s. Every version of scenario was repeated in the configuration: (A) without Message Broker and (B) with Message Broker.

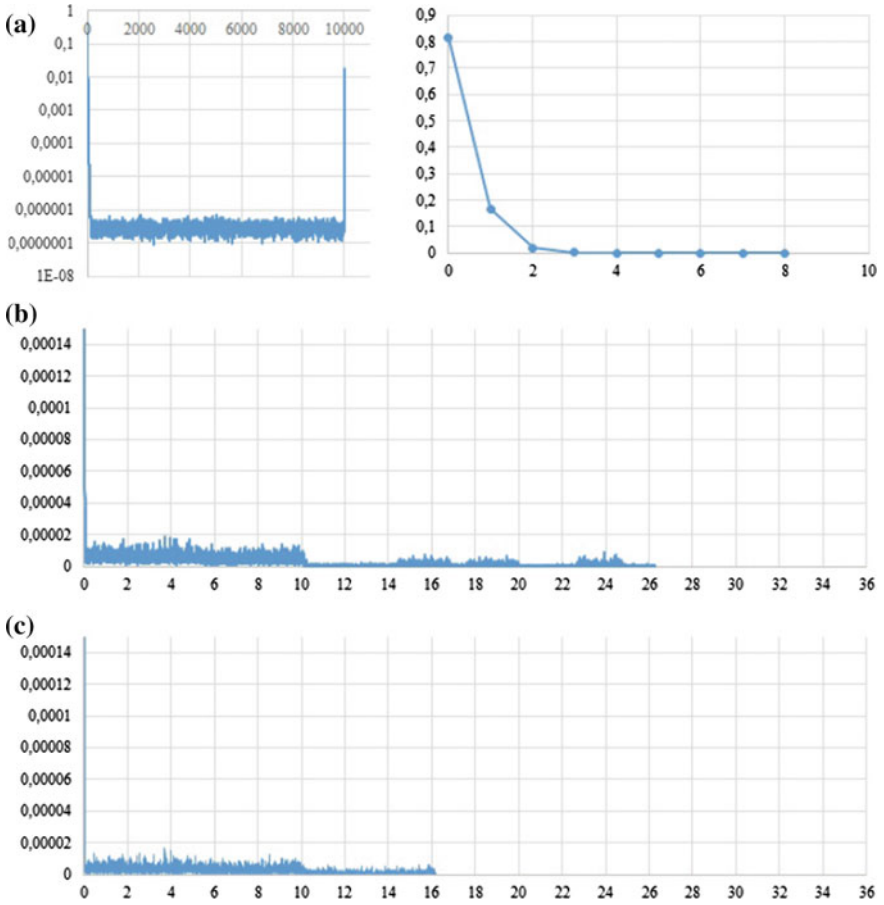


Fig. 3 Simulation results for scen. 1A. **a** Queue length distribution for GW0 (*left*) and LB (*right*). **b** Distribution of AP roundtrip-time. **c** Distribution of AP life-time

Particular charts show: (Fig. 3) the distribution of queue length in GW0 (thanks to load balancing, results for GW0 to GW3 for each scenario are indistinguishable); (Fig. 3) the distribution of AP life-time, defined as time from sending the packet from FS3 to receiving the packet by device; (Fig. 3) the distribution of AP packet round-trip time, defined as time from sending the packet from FS3 server to receiving acknowledgement back by FS3 distribution of queue length in load balancer (LB); for each scenario 1,2 and 3 in version A (without MB) and version B (with MB). In addition, in Table 5 information on packet loss probability is presented in reference to the GWs and FS1 server. Throughput information for GWs is presented, too. Packet life-time can be interpreted as real time from sending the order to its execu-

Table 5 GW throughput and packet loss probability for GW ans FS1

Scenario	1A	1B	2A	2B	3A	3B
GW throughput [pkt/s]	2804.67	2804.32	2847.04	2858.82	2915.54	2942.46
GW loss probability	0.0077718	0.012488	0.0092372	0.0089121	0.008479	0.010940
FS1 loss probability	0.161917	0	0.16191	0	0.162171	0

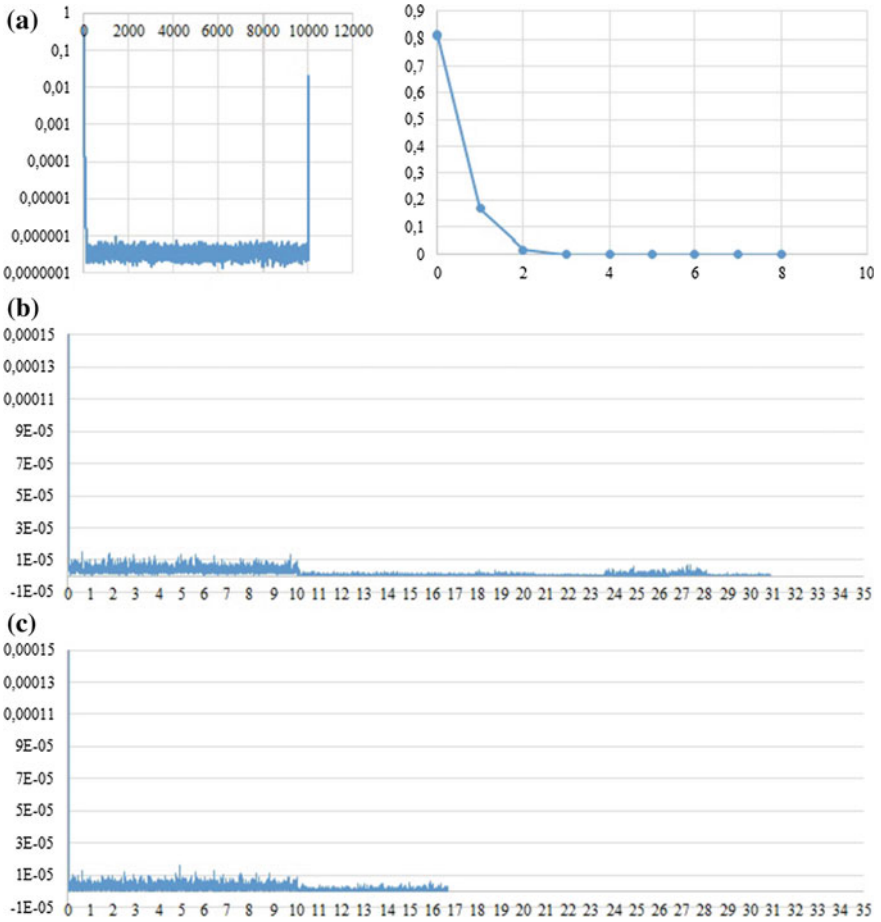


Fig. 4 Simulation results for scen. 1B. **a** Queue length distribution for GW0 (*left*) and LB (*right*). **b** Distribution of AP roundtrip-time. **c** Distribution of AP life-time

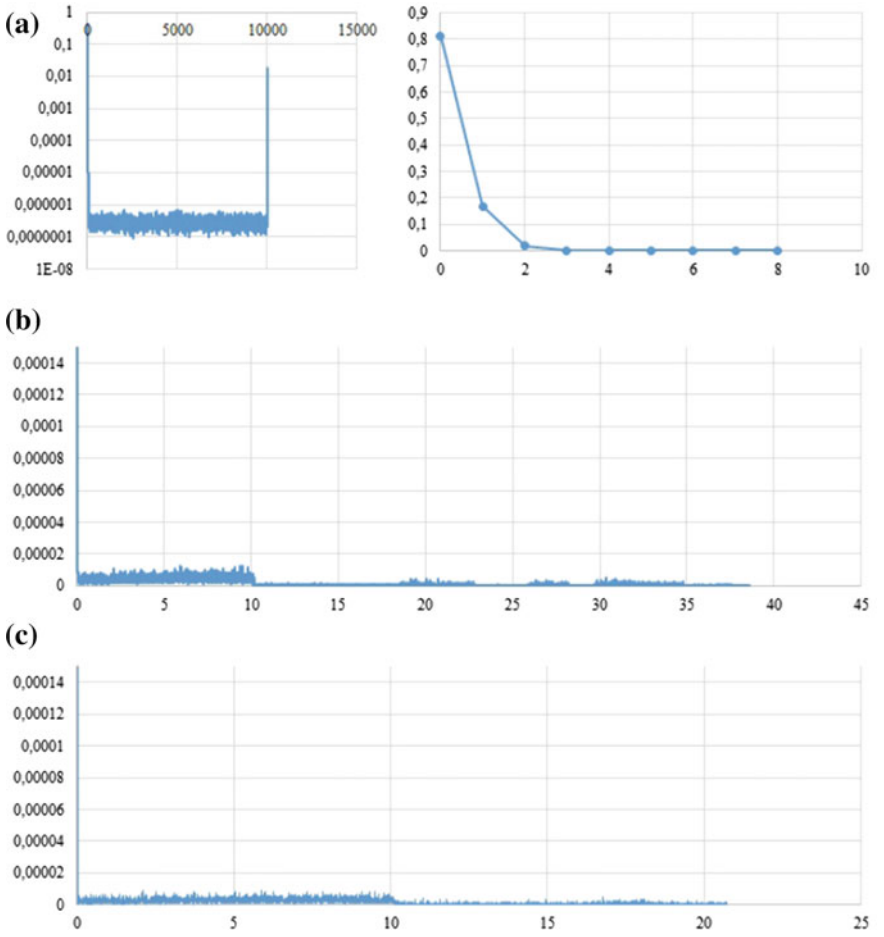


Fig. 5 Simulation results for scen. 2A. **a** Queue length distribution for GW0 (*left*) and LB (*right*). **b** Distribution of AP roundtrip-time. **c** Distribution of AP life-time

tion. Round-trip time is, in turn, time from sending the order to obtaining acknowledgement of its execution (Fig. 4, 5, 6, 7 and 8).

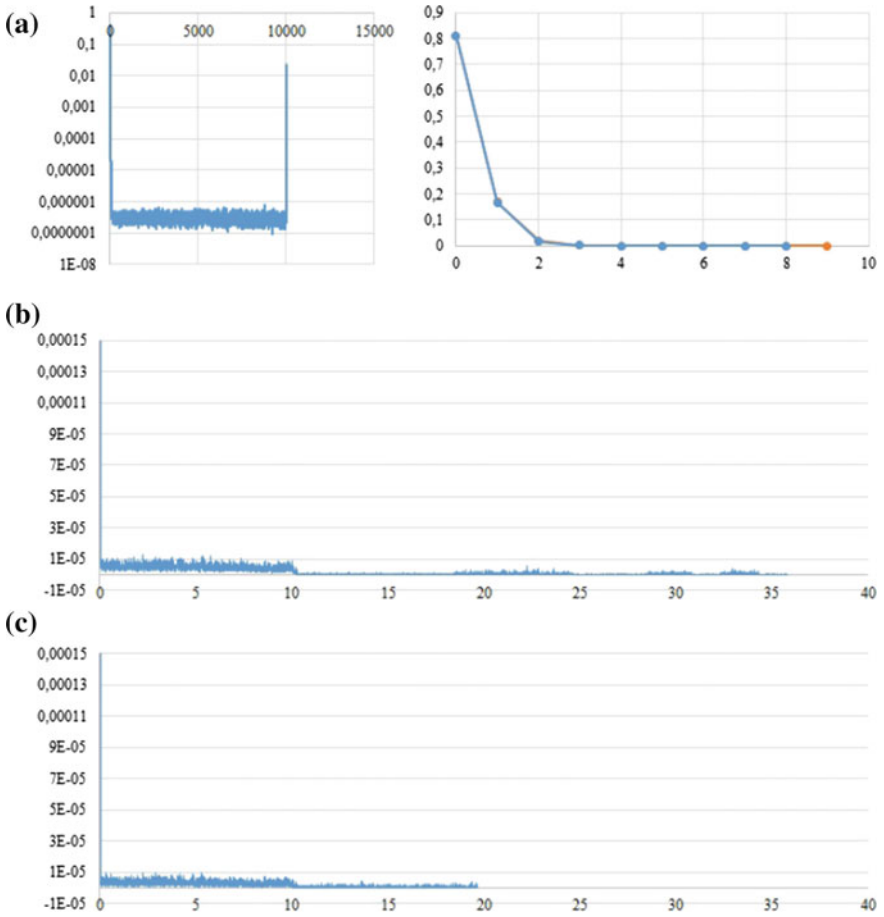


Fig. 6 Simulation results for scen. 2B. **a** Queue length distribution for GW0 (*left*) and LB (*right*). **b** Distribution of AP roundtrip-time. **c** Distribution of AP life-time

4 Conclusions and Practical Application of the Results

The analysis presented in the paper was focused on performance comparison of two variants of the architecture of distributed management system, however the results of the analysis show no significant performance difference between them. Both the architecture with centralised buffering in the message broker and the architecture

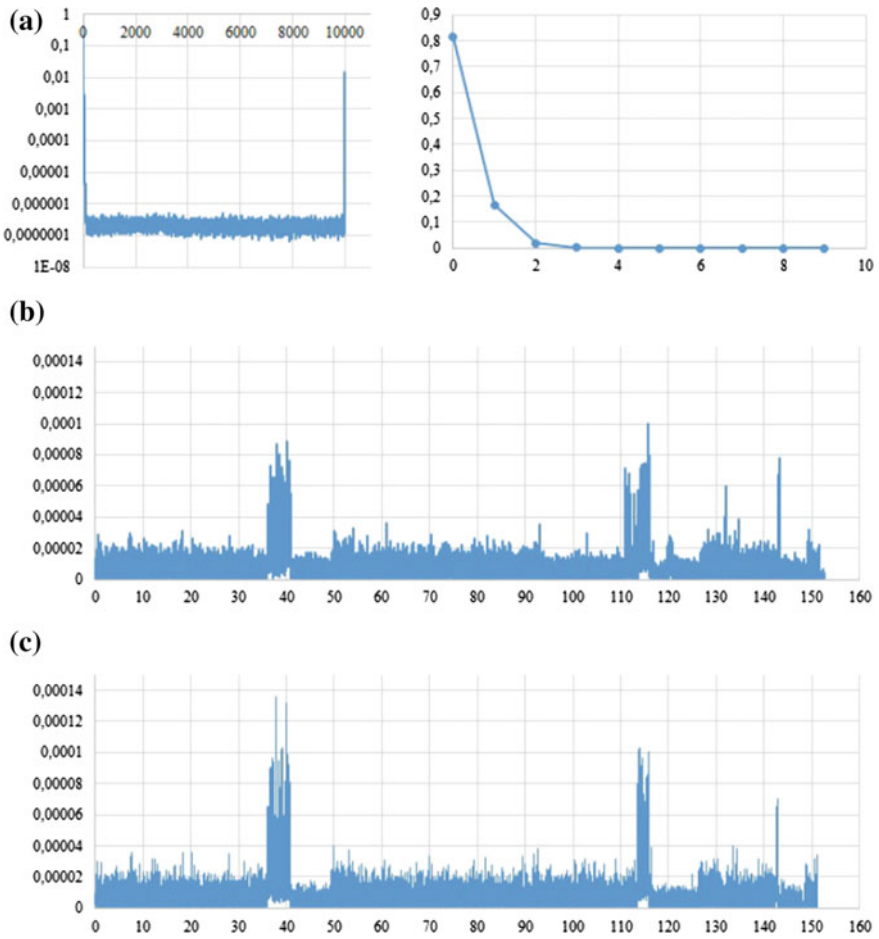


Fig. 7 Simulation results for scen. 3A. **a** Queue length distribution for GW0 (*left*) and LB (*right*). **b** Distribution of AP roundtrip-time. **c** Distribution of AP life-time

with distributed buffers may be used to gather the large data in the IoT network management system, serving millions of devices, as efficiency of the system is affected very little by the presence of MB. Packet loss probability in FSs goes to 0 in the system with MB, at the cost of higher loss ratio in GWs—but another GW can be added, and communication within the management system with centralised buffering is more reliable. It was necessary to take into account other factors affecting the quality of the created system. The architecture with central buffering is simpler in

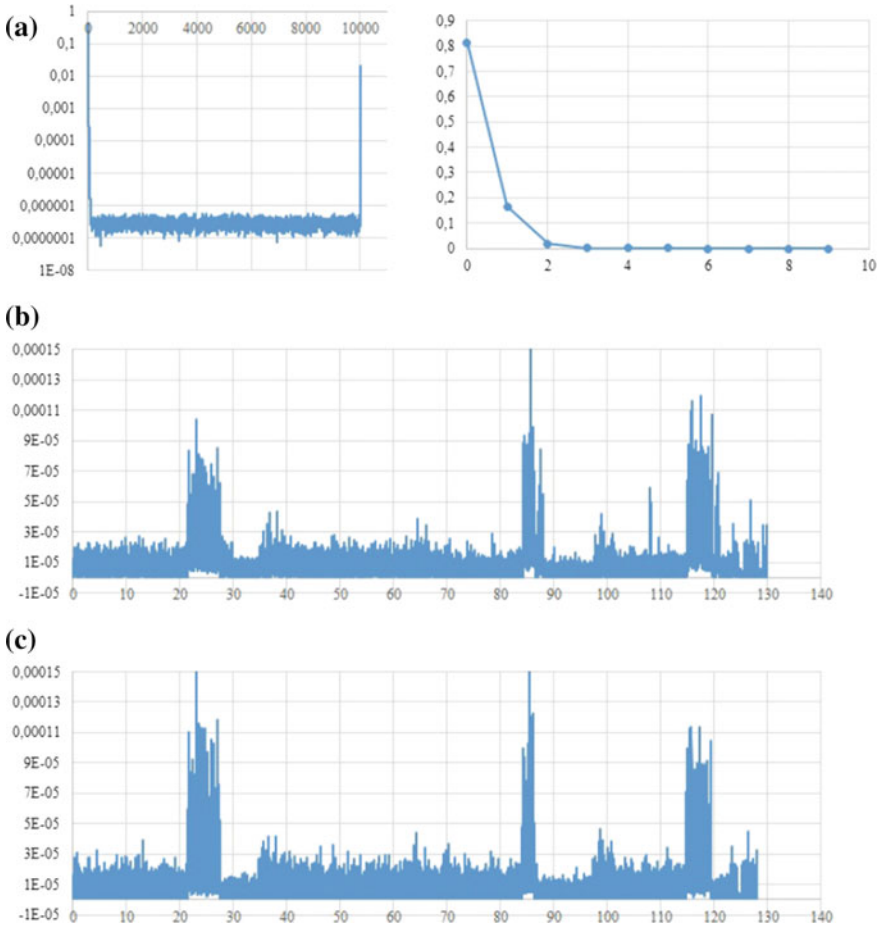


Fig. 8 Simulation results for scen. 3B. **a** Queue length distribution for GW0 (*left*) and LB (*right*). **b** Distribution of AP roundtrip-time. **c** Distribution of AP life-time

implementation, resulting in a system which is less buggy and can be build quicker. MOM systems became more and more popular in systems based on the publish-subscribe paradigm. The results of the analysis also demonstrate that the Markov chains are well suited to evaluating the performance of complicated modern network systems, providing results matching the discrete event simulation model.

References

1. Fox, G.C., Kamburugamuve, S., Hartman, R.D.: Architecture and measured characteristics of a cloud based internet of things. In: 2012 International Conference on Collaboration Technologies and Systems (CTS), pp. 6–12. IEEE (2012)
2. Proximetry, Inc. - home page. http://proximetry.com/
3. OASIS Advanced Message Queuing Protocol (AMQP) Version 1.0. OASIS Standard. http://docs.oasis-open.org/amqp/core/v1.0/amqp-core-complete-v1.0.pdf, Oct 2012
4. MSDN Library, Message Queuing (MSMQ). https://msdn.microsoft.com/en-us/library/ms711472(v=vs.85).aspx. Accessed 29 Dec 2015
5. Oracle, Java Message Service Concepts. http://docs.oracle.com/javaee/6/tutorial/doc/bncdq.html, Jan 2013. Accessed 29 Dec 2015
6. The Apache Software Foundation, ActiveMQ. http://activemq.apache.org/. Accessed 04 Jan 2016
7. Rostański, M., Grochla, K., Seman, A.: Evaluation of highly available and fault-tolerant middleware clustered architectures using RabbitMQ. In: 2014 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 879–884. IEEE (2014)
8. Desai, P., Panse, A., Jadhav, M., Gavhane, A., Patwardhan, A.: Fiesta: parallelism for data collection and intelligent inference in a distributed heterogeneous environment. In: Proceedings of the 2011 UKSim 5th European Symposium on Computer Modeling and Simulation, EMS '11, pp. 237–240. IEEE Computer Society, Washington, DC, USA (2011)
9. Jia, Y.: Resilient and Efficient Delivery over Message Oriented Middleware. Ph.D. thesis, School of Electronic Engineering and Computer Science Queen Mary, University of London, Sept 2014
10. Psiuk, M., Żmuda, D., Zieliński, K.: Distributed OSGi built over message-oriented middleware. Softw. Pract. Exp. **43**(1), 1–31 (2013)
11. Stewart, W.J.: Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling. Princeton University Press, Oxford, Princeton (N.J.) (2009)
12. Pecka, P., Deorowicz, S., Nowak, M.: Efficient representation of transition matrix in the Markov process modeling of computer networks. In: CzachA3rski, T., Kozielski, S., Stanczyk, U. (eds.) Man-Machine Interactions 2. Advances in Intelligent and Soft Computing, vol. 103, pp. 457–464. Springer, Berlin, Heidelberg (2011). doi:10.1007/978-3-642-23169-8_49

Communication Approach in Distributed Systems on .NET Platform

Aneta Poniszewska-Maranda and Piotr Wasilewski

Abstract Although the history of distributed applications goes back to the 60s of the last century, the tremendous growth of opportunities faced by software developers has been in recent years. With the continuous increasing of access to the Internet, both in the traditional way using a desktop computer, and more often used mobile devices such as phones or tablets, the demand for providing more complex distributed systems increases. To meet their requirements it is necessary to introduce new solutions that will be able to handle a very large number of users from around the world. The paper presents the analysis of different ways of communication between distributed applications in .NET environment.

Keywords Distributed systems · Communication between distributed applications · .NET platform

1 Introduction

Distributed applications date back to its history until the 60s of the last century. Operating systems as the first were able to perform a variety of calculations at the same time by switching between threads. The deployment of Ethernet in the 70s led to the rapid development of distributed systems. Using the network for communication between applications located within different computers is one of the key issues in distributed systems running smoothly. The challenges faced by architects of distributed application are as follows [1, 2]:

A. Poniszewska-Maranda (✉) · P. Wasilewski
Institute of Information Technology, Lodz University of Technology, Łódź, Poland
e-mail: aneta.poniszewska-maranda@p.lodz.pl

P. Wasilewski
e-mail: wasilewski.pio@gmail.com

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_34

- *heterogeneity*—diversity of equipment, protocols, programming languages, architectures and operating systems,
- *openness*—gives the possibility of extending the existing distributed system both in terms of hardware, for example by including new computers to increase the computing power of the whole system and in terms of software, i.e. providing the interface for system developers,
- *security*—it is more and more important in software development because of growing popularity of systems, i.e. safety of messages sent between computers belonging to the distributed system is based not only on encryption of their content, but also on clear statement, if someone did not intercept and modify the messages or send specially crafted message (*spoofing*),
- *scalability*—ability to provide a corresponding increase in system capabilities, along with an increase in the number of users using it,
- *errors handling*—essential in each application but the *complexity* of distributed systems hinders the proper errors handling—there are often a number of different components, deployed on computers in multiple locations, which makes the error detection also difficult,
- *concurrency of data access* is very important for correct application operation—in many distributed applications, an access to data must be implemented in appropriate manner, so as to avoid the deadlocks and to provide the appropriate order for access to record the new data,
- *transparency* of the system is to create it in such a way that the user does not aware that the application he is using, is only a part of a larger system deployed on other computers.

With the enormous progress of computer science and a wide access to Internet, the development of distributed applications has changed. Currently, there are a lot of libraries and programming languages, specifically designed to support the creation of this type of applications/systems. They allow the developers to focus on meeting the business and functional requirements of applications, instead of dealing with low-level aspects of the communication between distributed applications.

The problem presented in the paper concerns the analysis and comparison of different ways of communication between distributed applications in .NET environment. The main emphasis has been placed on selected concepts of this platform. The presented paper is structured as follows: Sect. 2 gives the outline of communication models for distributed applications and distributed communication in .NET environment. Section 3 deals with approach of communication of distributed applications based on .NET platform, while Sect. 4 describes different tests of chosen libraries for solving the communication problem between distributed applications.

2 Communication of Distributed Applications

Continuous development of software and Internet has meant that nowadays the distributed solutions can achieve things previously impossible. Banking systems, online reservations, social networks benefit depending on the needs of different communication models outlined below in this section.

2.1 Communication Models of Distributed Applications

The models of communications between distributed applications can be regarded as [1, 2]:

- *interprocess communication*—low-level solution, which consists of sending simple messages between processes, for example by TCP or UDP,
- *indirect communication*—loose connection between sender and receiver through an intermediary—a broker (messages queue, group communication), characterized by the separability of space (*space uncoupling*) or separability of time (*time uncoupling*),
- *remote call*—exchange of messages involving the remote execution of procedure or method as it was located locally on the computer.

The most commonly used solution is *remote procedure call, RPC* [3]. According to the creation of distributed applications, communication should be transparent. The user executing the procedure locally on his computer, should not be aware that it is pursued on remote computer.

The remote procedure call fills this assumption as follows: when a client (local machine) wants to call a procedure, the application is halted and the message with parameters is passed to the server (remote machine) that performs the procedure. Then, when it finishes the processing it sends the return message, containing the results of called procedure (Fig. 1).

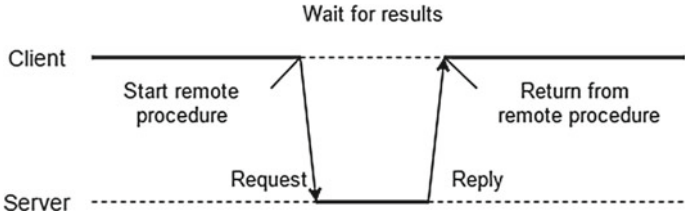


Fig. 1 Schema presenting the remote procedure call [4]

Indirect communication has been defined as communication between applications using an intermediary, without a direct connection between the sender and receiver [1]. It results in two basic features of such solution:

- space uncoupling—lack of mutual knowledge about identity of sender and receiver; it allows to modify or update any communication participants without much interference in the system,
- time uncoupling—describes the lifetime independence of communicating objects; it means that they do not need to exist at the same time to communicate with each other.

2.2 Distributed Communication in .NET Environment

To implement a distributed system the ready-made libraries can be used to allow the easy communication between applications located on different computers. The solutions available in .NET platform are as follows:

- *.NET Remoting*,
- *ASP.NET web services*,
- *Windows Communication Foundation*.

.NET Remoting [5] allows the communication between *application domains*. One or more domains can be created within a single process. It allows to create a separate virtual memory and various security rules for each of them. The use of .NET Remoting for communication between domains can be done within a single process, between processes on the same computer or between processes on different computers. It is based on distributed object model. It is possibly to get an access to objects created in another application domain.

ASP.NET is the development platform that allows to create the web applications and web services. *Web service* is defined as the application logic which is available for other programs via standard network protocols independently of platform or system on which it operates [6, 7]. The key to the success of web services is the use of existing solutions and technologies in order to achieve the desired features, such as connecting of applications written in different programming languages. This is a very important aspect from the point of view of the users of existing systems written on different operating systems or platforms. Using web services we can combine both new applications with new, as well as with those that already exist.

Windows Communication Foundation, WCF is a library created to facilitate the design and programming of distributed applications. *WCF* combines the advantages of all previous technologies: *.NET Remoting*, *Microsoft Message Queuing*, *COM+, Enterprise Services* and network services [8, 9].

3 Communication of Distributed Applications Based on .NET Platform

The aim of this paper is to present the analysis and comparison of different ways of communication between distributed applications in .NET environment. The following section presents a skeleton of application, allowing the comparison of performance of libraries described in Sect. 2.2.

Solutions based on .NET platform operate using the *client-server* communication model. Therefore, the created solution of communication problem was split into four applications (Fig. 2). Three of them represent the servers handling the user requests and one is a client application sending the requests.

The comparison of individual solutions for communication between distributed applications based on .NET platform relies on comparison of total time of procedure execution on the server. In order to obtain the reliable results, the *test scenarios* were defined that will allow to carry out the same measurements for all selected libraries:

- sending of *built-in type variable*, e.g. integer or text character, and receiving the value returned by the server,
- defining of *object type* whose fields consist only of built-in types—the variable created in such way is sent to the server as a parameter, and then received back by the client,
- creating of *complex object type*, that contains other object, fields and tables,
- sending and receiving of *array* consisting of built-in variables,
- using of *array* containing the complex object types.

For comparison of applied tools the algorithm presented in Fig. 3 was used. The implementation of client application with the use of libraries based on .NET platform requires the creation of a server, which processes the data according to transmitted parameters. The attempt to connect to the server is taken every time the application is

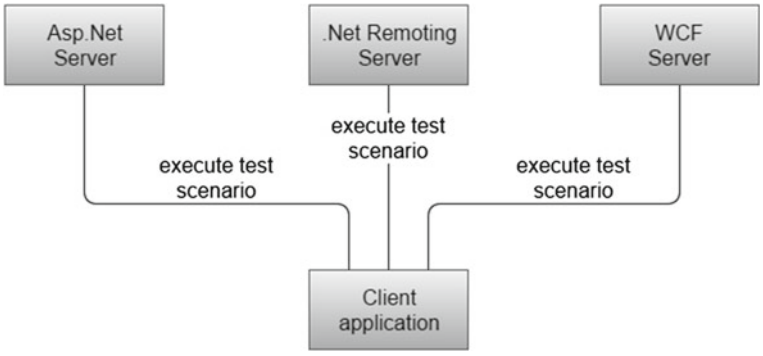


Fig. 2 Server and client software required to solve the problem of communication between applications

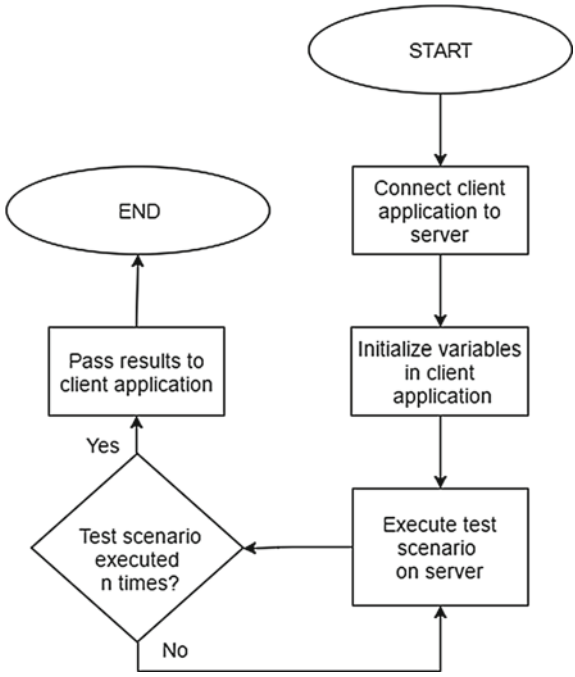


Fig. 3 Algorithm measuring the execution time of test scenario in proposed solution

turn on. If it is not completed successfully—the user should be notified in appropriate manner.

Before calling the required procedure on the server, the time of its execution should be started to measure. It will allow for later comparison of libraries in terms of their performance. Every test scenario should be made at least 500 times to obtain the reliable results. The final result should be presented in the form of arithmetic average of the total execution time of all calls. The final stage of the algorithm is to transfer and present the results to the user.

Libraries selected for the realization of presented solution are based on .NET platform. For this reason the test scenarios will review only the intermediary (Fig. 4), but not the performance of programming language. The comparison of libraries requires the use of identical implementation of scenarios for all libraries.

The research also check the impact of changes of basic configuration elements of the libraries to facilitate the communication between distributed applications:

- format of sending data (SOAP, binary) to which the parameters are serialized/deserialized, names of methods and their results,
- different transport protocols (TCP, HTTP), which also affect the time of procedure execution.

Basing on the characteristics of selected libraries and the description of existing implementations of distributed applications, the functional requirements of the client software are specified:

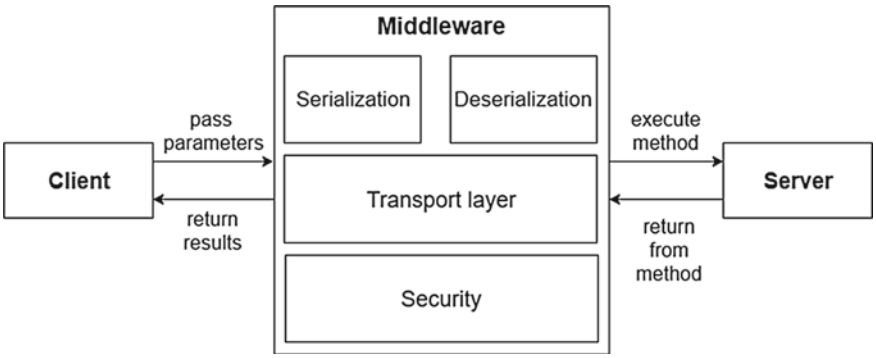


Fig. 4 Schema presenting the intermediary (broker) between client and server

- measurement of execution time of method chosen by a user on the server,
- connecting to the server,
- changing of intermediary configuration (form of data encryption, transport protocol),
- possibility of changing of test scenarios,
- calling a method that supports the selected scenario.

Functional requirements of the server supporting the queries of client software includes: processing of methods selected by the user, return of obtained results and listening for new connections.

An important advantage of proposed approach is to use the known solutions based on .NET platform. Accurate documentation of *.NET Remoting*, *WCF* and *ASP.NET web services* allows to check all necessary implementation details. The large community of programmers working for many years with these libraries also helps in such solutions. Another advantage is the ability to immediately work on logical application layer because the provided tools with the appropriate configuration will deal alone with aspects related to the communication between distributed applications.

4 Tests of Libraries for Communication Problem Between Distributed Applications

To carry out the tests enabling the comparisons of libraries for creating of distributed applications, the solution discussed in previous section was implemented. The created software allows to measure the execution time of remote method on the server, executed by the user from client application level. Measurements made using a high-resolution timer, allow a fair comparison of performance of the following libraries: *.NET Remoting*, *ASP.NET web services* and *Windows Communication Foundation*.

The client software has been created to verify how much time the call and execution of the remote method on the server that is located in a different process than the client will take. The research is based on a simple method call, because its purpose is to check the library elements responsible for communication between applications.

Performance tests of used solution were carried out using different configurations of the libraries. Both *Windows Communication Foundation* and *.NET Remoting* support the following protocols and data encoding formats:

- *HTTP* and *SOAP* messages,
- *HTTP* and *binary* encoding,
- *TCP* and *SOAP* messages,
- *TCP* and binary encoding,

The above combinations of transport protocols and data encoding formats were selected for the experiments because they are common for two test libraries. The exception is the implementation of a network service based on *ASP.NET*—it supports only the connection with the use of *HTTP* and *SOAP* messages.

The first conducted experiment was to verify the performance of libraries in the case of transfer a primitive variable as a parameter of remote procedure. Table 1 presents the results obtained from local and remote tests. The results are correct with the expectations. Protocol *TCP*, either in conjunction with a binary coding and *SOAP* messages is faster then *HTTP*. The fastest library in case of primitive variables is *.NET Remoting*.

Second test is the performance of the library in case of transfer a simple class as a parameter. Declared object consists of two integers, so do not differ too much from a variable of embedded type. Table 2 confirms this assumption, because the results do not differ too much from the previous test.

The next test is to check the impact of data encoding format and transport protocol on the time of execution of remote method which parameter is the complex class. It consists of two simple objects, list and array, containing ten thousand elements each. The results are presented in Table 3. Only in case of more complex objects

Table 1 Timing of remote method on the server in scenario where the transmitted parameter is primitive variable

	HTTP/SOAP (ms)	HTTP/Binary (ms)	TCP/SOAP (ms)	TCP/Binary (ms)
Methods invoked locally				
WCF	1.02	0.94	0.78	0.74
ASP.NET	2.72	–	–	–
Remoting	0.39	0.22	0.34	0.17
Methods invoked remotely				
WCF	21.95	20.96	2.81	2.59
ASP.NET	31.12	–	–	–
Remoting	20.59	21.95	23.47	20.26

Table 2 Timing of remote method on the server in scenario where the transmitted parameter is simple object variable

	HTTP/SOAP (ms)	HTTP/Binary (ms)	TCP/SOAP (ms)	TCP/Binary (ms)
Methods invoked locally				
WCF	1.24	1.07	0.80	0.75
ASP.NET	2.91	–	–	–
Remoting	0.35	0.20	0.33	0.17
Methods invoked remotely				
WCF	22.73	20.24	3.27	2.88
ASP.NET	34.78	–	–	–
Remoting	25.84	24.49	26.19	22.27

Table 3 Timing of remote method on the server in scenario where the transmitted parameter is complex object variable

	HTTP/SOAP (ms)	HTTP/Binary (ms)	TCP/SOAP (ms)	TCP/Binary (ms)
Methods invoked locally				
WCF	2.12	1.97	1.84	1.73
ASP.NET	630	–	–	–
Remoting	6.93	5.97	6.73	5.34
Methods invoked remotely				
WCF	21.04	20.62	3.28	2.84
ASP.NET	1131.12	–	–	–
Remoting	31.94	27.15	29.38	26.54

WCF begins to achieve the better results. Significant reduction in performance can be observed during the tests carried out using *ASP.NET*.

Results from Table 4 presents the speed of method execution which parameter and result is an array consisting of ten thousand elements of primitive type. In case of array, there is much higher update of method execution speed when the data encoding format is changing. And the protocol does not play any great role. This is due to the construction of *SOAP* message which for each of ten thousand elements of an array creates a new *XML* item and in case of binary encoding it is a single byte stream.

Similar results as in the previous test, are also visible in Table 5, presenting the execution time of remote method on the server for the test scenario, where the transmitted parameter is an array of simple objects (ten thousand simple objects in performed tests). The reason is a simple object consists of two primitive variables, so the serialization method differs only in that the element that wraps them is a class definition.

All above results indicate that there is no universal solution of communication problem between distributed applications. If it operates only on primitive variables, the use of *.NET Remoting* can be considered. This library achieves the best results

Table 4 Timing of remote method on the server in scenario where the transmitted parameter is array consisting of primitive variables

	HTTP/SOAP (ms)	HTTP/Binary (ms)	TCP/SOAP (ms)	TCP/Binary (ms)
Methods invoked locally				
WCF	22.03	1.33	21.34	1.08
ASP.NET	31	–	–	–
Remoting	26.60	0.63	26.25	0.55
Methods invoked remotely				
WCF	286.51	24.21	256.12	17.58
ASP.NET	378.12	–	–	–
Remoting	310.45	295.51	320.15	290.86

Table 5 Timing of remote method on the server in scenario where the transmitted parameter is array consisting of simple objects

	HTTP/SOAP (ms)	HTTP/Binary (ms)	TCP/SOAP (ms)	TCP/Binary (ms)
Methods invoked locally				
WCF	15.76	12.67	14.57	12.01
ASP.NET	33	–	–	–
Remoting	24.51	0.48	25.48	0.41
Methods invoked remotely				
WCF	121.50	95.79	119.67	37.09
ASP.NET	234.12	–	–	–
Remoting	183.09	150.73	175.48	120.90

using *TCP* transport protocol and binary format of data encoding. The disadvantage of such solution is the communication only between the applications based on .NET platform. The use of *WCF* gives greater flexibility. Using of *HTTP* protocol and *SOAP* messages the satisfactory performance can be obtained while maintaining the independence of a particular platform. The lowest efficiency presents the *ASP.NET* network service that turned out to be the slowest in all carried out tests.

5 Conclusions

Communication between different programs of a large system is a key aspect of distributed applications development. To create a high-quality software that meets the latest standards, the use of middleware in exchange of messages between applications residing on different computers is recommended. Solutions based on .NET platform provide the ability to communicate between programs located on different

computers in safety manner, in accordance with generally accepted standards, so the heterogeneity existing in all Internet networks does not constitute any barrier.

Distributed applications offer the great opportunities that are not possible to obtain on a single machine. They include applications based on voluntary computing, social networks and *peer-to-peer* platforms for the exchange of files.

The paper presented the analysis and performance tests of different libraries included in .NET development platform, used for communication between distributed applications. The solutions are: *.NET Remoting*, *Windows Communication Foundation* and *ASP.NET web services*. This choice was motivated by their similarity in the aspects of communication—each used solution provides the support for *HTTP* transport protocol and format of data encoding in form of *SOAP* messages. As it was expected, there is no ideal solution that can be used in all possible situations. The creation of distributed application, which will meet the requirements of users and its architects, requires the careful analysis of possible solutions. It will allow the selection of appropriate library to the identified criteria.

References

1. Coulouris, G.F., Dollimore, J., Kindberg, T., Blair, G.: Distributed Systems: Concepts and Design, 5th edn. (2012)
2. Birman, K.P.: Reliable Distributed Systems, Technologies, Web Services and Applications (2005)
3. Birrel, A.D., Nelson, B.J.: Implementing remote procedure calls. ACM Trans. Comput. Syst. **2** (1984)
4. Tanenbaum, A.S., Van Steen, M.: Distributed Systems Principles and Paradigms, 2nd edn. (2007)
5. Nagel, C., Evjen, B., Glynn, J., Watson, K., Skinner, M.: Professional C# 4 and .NET 4 (2010)
6. Basiura, R., Batongbacal, M., Bohling, B., Clark, M., Eide, A., Eisenberg, R., Hoffman, K., Loesgen, B., Reynolds, C., Sempf, B., Sivakumar, S.: Professional ASP.NET Web Services. Wrox Press (2001)
7. Uurlu, A., Zeitler, A., Kheyrollahi, A.: Pro ASP.NET Web API: HTTP Web Services in ASP.NET. Apress (2013)
8. Klein, S.: Professional WCF Programming: .NET Development with the Windows Communication Foundation. Wiley Publishing (2007)
9. Lowy, J.: Programming WCF Services, 3rd edn. O'Reilly Media (2010)

Personalisation of Learning Process in Intelligent Tutoring Systems Using Behavioural Measures

Piotr Chynał, Adrianna Kozierkiewicz-Hetmańska and Marcin Pietranik

Abstract The main goal of an intelligent tutoring system is to provide learning materials suitable for students' needs and preferences. Observations and analysis of students' behaviour and interactions with the intelligent tutoring system are crucial to determine and, if necessary, modify the learning scenario. A properly designed user's model influences the effectiveness of those methods and, in consequence, the whole learning process. This paper is devoted to propose a content of the student's profile that includes behavioural measures.

Keywords EEG · Emotion recognition · Learner profile · Eyetracking · Intelligent tutoring system

1 Introduction

The idea of continues education was first mentioned by Basil Yeaxlee in 1929 [31]. In his first book, the author has described the education as a part of everyone's life. The process of continuous improvement and development of knowledge and skills is not only limited to children and young people who attend the schools, but it is also for adults who want to improve their skills, develop new one or gain a new profession.

The mentioned goal can be achieved by learning with help of intelligent tutoring system. Its main objective will be to improve the efficiency of the education process. In this paper the model of an intelligent e-learning system which takes into

P. Chynał (✉) · A. Kozierkiewicz-Hetmańska (✉) · M. Pietranik (✉)
Wrocław University of Science and Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: piotr.chynal@pwr.edu.pl

A. Kozierkiewicz-Hetmańska
e-mail: adrianna.kozierkiewicz@pwr.edu.pl

M. Pietranik
e-mail: marcin.pietranik@pwr.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_35

account various biometric measures is presented. The idea of the system is based on the assumption that similar students learn in the same or a very similar way. The adoption of that assumption entails that the learning is a process consisting of several steps.

The first step is gathering the most relevant data that describes the student. During the registration the intelligent tutoring system collects information about the student. This data will be collected not only using direct surveys filled by the student, but also using psychological questionnaires and various devices recording the behaviour and the emotions of the user. In the next step the student is classified to a group of similar users and starts learning. Users and usage data stored in the system are the basis for personalization and recommendation of learning process.

Due to the fact that students differ between each other, therefore the structure and organization of educational material can be very convenient for one, and difficult to comprehend for another. The material presentation, from general to specific, will be beneficial for people with holistic learning style. The opposite situation occurs in the case of analytical style [19]. Online tools provide some additional capabilities that allow linear, hierarchical and relational arrangement of the material. Their impact on academic performance is also dependent on the styles of learning [6]. Personalization of learning can raise the academic performance up to approximately 7–8 % [13].

We assume that an initial scenario for the new user is selected from the completed learning scenario of students, who belong to the same group. Similarity is based on all collected data. By "the learning scenario" we will call the educational material presented in a certain, fixed order and in an appropriate form. The choice of the scenario in the initial, basic form will be proposed to the user after the registration process and is the key element of the designed system.

After generating the appropriate scenario, the student can begin the actual learning process. However, in many situations the student's learning results could be insufficient. Then, it is a signal to the system that the proposed learning scenario is not appropriate for the student who should be offered a modified learning scenario.

Methods of modifications of the learning scenario incorporate information stored in the system that concern the completed scenarios, obtained results and the current characteristics of the user all of which were collected during the student's interaction with the system. Thanks to collecting of many behavioural measures system will be able to identify the reason of student's mistakes and propose the appropriate modification of learning scenario. For example, the eye-tracking could be helpful with assessment of student's concentration on learning material. The controlling of the user's pulse could be important for estimation of level of anxiety and stress. The wide knowledge about student's characteristic plays the important role in personalization of the learning process. The general idea of an intelligent tutoring system is shown in Fig. 1.

The use of the intelligent e-learning system brings many benefits. Research confirms that the use of profiling students and selection of scenarios based on their predisposition increases the motivation to learn and learning outcomes [1].

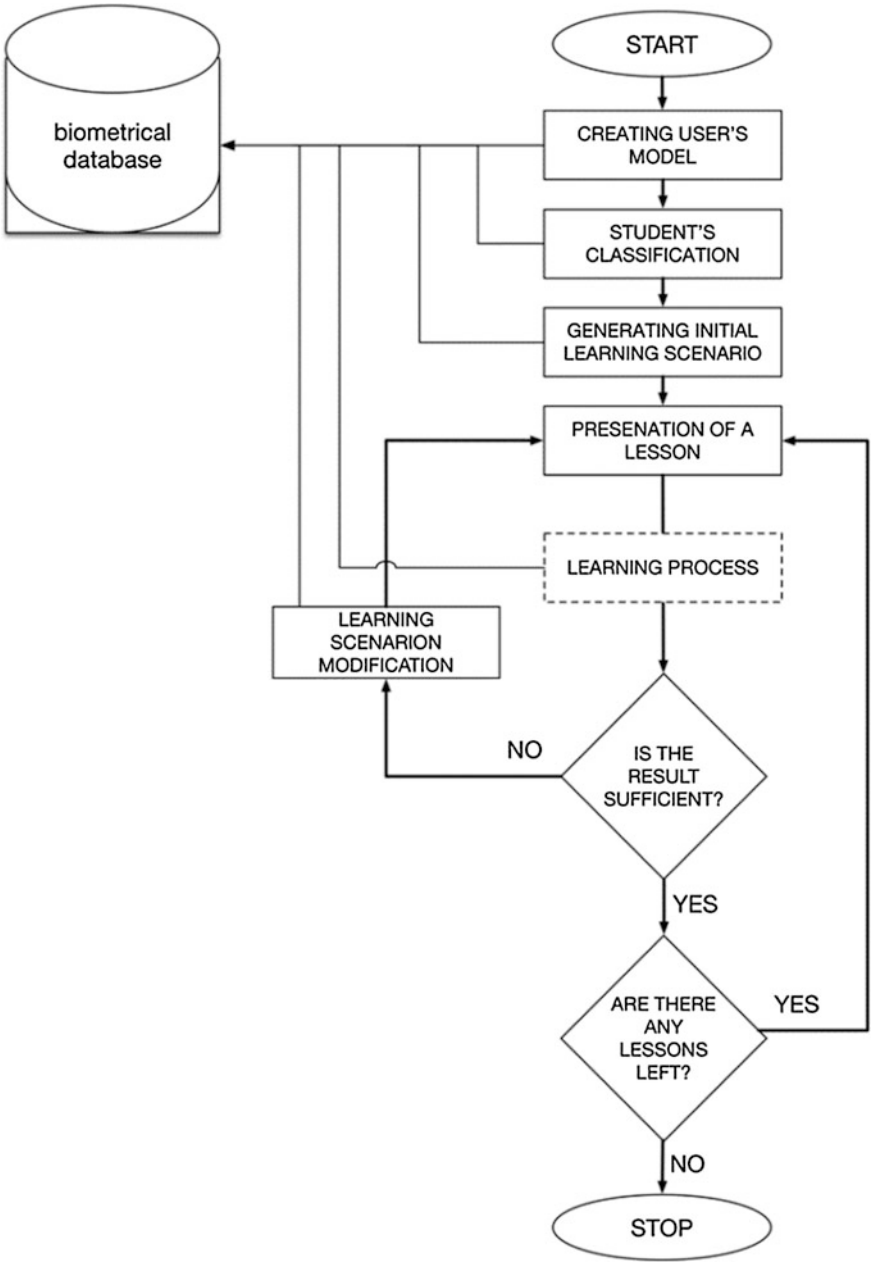


Fig. 1 The general idea of an intelligent tutoring system

Therefore, the aim of this research is to improve the efficiency of learning by developing learner model containing biometric data. Such, well designed, model is very useful for adjusting material recommendations to the preferences, needs, learning styles, abilities [14], as well as the current emotional state of students.

The structure of this paper is as follows, the second chapter presents theoretical information about intelligent tutoring systems and the current state of art in this field. Third chapter introduces a research problem and describes how we are going solve it. Last chapter concludes the whole article.

2 Intelligent Tutoring Systems State of the Art

The contents of the user's model have a major impact on the efficiency of intelligent tutoring systems. If the system has sufficient information about a student, it can better control the process of education, which in turn improves the quality of teaching [12].

The profile of the student consists of two types of data: user's data (referring to the personal characteristics of the user, such as demographic data, knowledge, interests, preferences, intentions) and usage data (history of student's interaction with the system). The former includes personal data, which only allow to identify the student in the system (e.g. username, e-mail etc.). Virtually every system stores such data in users' profiles which are frequently supplemented by attributes such as age, education level, gender, that may be helpful later on in the personalization process [12].

Another type of data are descriptions of learning styles. They describe how the acquisition, processing and receiving information from the surrounding world affects the student- intelligent learning systems should create a learning environment that the user prefers. In the literature, there are many well-known theories of learning styles and some of them are successfully used in e-learning systems.

For example, in the INSPIRE [7] system the Honey and Mumford questionnaire was used. The authors of this questionnaire identified four types of students: activists (that prefer learning by experience), reflection (preferring to watch a material before taking action), theorists (that prefer the adaptation and integration of all observations in the form of theory or framework) and the pragmatists (looking for new ideas that may be implemented practically). Model of Felder and Salomon were applied in [16]. Others systems storing information about student's learning style were described in [3, 28].

The last type of data typically stored in a profile are usage data. These are all the information gathered during user interaction with the intelligent tutoring system. Depending adopted by the designer of the system assumptions, these data may be more or less detailed. The most common user model includes the time spent on learning the particular lesson [27]. The intelligent tutoring systems also store the result obtained in the test concerning some selected batch of material [3]. An important role is played not only by the final result of the test, but also the number

of correct, incorrect and all of the student's answers [29]. The data are further analysed by the system and used to modify learning scenarios to propose appropriate repetitions. Many intelligent tutoring systems store also data concerning users' surveys, tests and evaluation results [29, 30].

None of the analysed systems neither in the generation of the initial scenario nor its subsequent adaptation to the user's requirements did not include biometric data collected in parallel to the learning process [26].

The utilization of information from eye-tracker, electroencephalograph, devices recording electrodermal activity and measuring person's pulse [17, 25] not only allows to customize adaptive and intelligent user interface of the intelligent tutoring system, but also provides means to dynamically modify learning scenarios. In literature we have not found a significant number of publications concerning the combination of these two issues [9]. Furthermore, analysed articles focus on the usability of the intelligent tutoring system itself [2]. Only one publication involved a comprehensive analysis of student behaviour involving their emotional state at the time of learning [21]. However, the project was based solely on the eye-tracking data and not on a holistic analysis of the biometric descriptions of students, their expressed emotions in text messages, their learning results and openly expressed preferences.

3 User's Profile Containing Biometrical Data

Our research focuses on developing the model of an intelligent learning system which takes into account various biometric indicators. Its main objective will be improving the efficiency of the education process. Increase of the efficiency of the learning process is enabled by modern methods of obtaining information about users. The system can collect not only basic information such as age, gender, skill level, and the order of the learned lessons. Today it is possible to perform detailed tracking of user behaviour, including tracking of eye movements using eye-tracking methods, pulse, temperature changes of the skin on users face etc.

The contents of user model has a major impact on the efficiency of education systems. If the system has sufficient information about the student, it enables better control of the process of education, which improves the quality of teaching. In the student profile two types of data are stored: user data (refers to personal characteristics of the user, such as demographic data, knowledge, interests, preferences, intentions) and usage data (history of students' interaction with the system). The content of the user profile, depending on the needs, may be more or less detailed. Larger number of data stored in the user model allows for better matching of material and learning environment to the preferences of the student, but is also related to e.g. the extension of the registration process where the user is forced to provide information about himself, fill in various questionnaires, or pre-tests examining the initial level knowledge of the student. The goal is to find significant features that describe the students that have an impact on the process of the

recommendation of teaching material. The intelligent educational system can distinguish several groups of students' features that the system should store.

These include demographic data, allowing to only distinguish students in the system, and the initial level of knowledge of the student. Another important element are the learning styles, which also imply the choice of components. Learning styles will be determined based on the model of Felder and Solomon, and evaluated using the Index of Learning Styles Questionnaire [11]. Additionally, based on student's profile, it should be possible to describe user abilities, interests or personal character traits. This type of content of learner model and analysis of its attributes was described in more details in [14].

However, the most important element of user profile will be the biometric data collected during students' interaction with the system. This will allow to assess the emotional state (anxiety levels, concentration, general emotional state, engagement etc.), of a student in the course of working with the education system and thus, enabling more effective personalization of the user's profile.

For collecting biometrical data, there are many devices that can be used. First of them is an eye-tracker, which allows to follow and record eye movements, and accordingly analyse on which elements of the presented content the student focused most of his attention [4]. Research conducted by eye-tracker provides extensive knowledge of the behaviour and unconscious preferences of the user. They allow to check if: the readable content is ignored, users focus their gaze on the most important topics, the presented information are selectively absorbed, there is no information overload, navigation is unreadable, and how long was the time of eyes' focus on specific content.

Software used in eyetracking provides many different graphical reports, illustrating the test results. The most popular are: heat maps which present places on which user vision was mostly centred and fixation maps that present the route of the users gaze, and also the points at which he looked for a longest time, along with information how much time it lasted.

Another device intended to be used in this research is an electroencephalograph (EEG) [25]. Low-cost example of such device is Emotiv Epoc, which can detect five different types of emotions (channels): instantaneous excitement, long term excitement, meditation, engagement and frustration [10]. The determined scores on each channel range from zero to one, where a higher channel score corresponds to a greater intensity of the emotion. The definition of the detected emotions is following [5]: *frustration* (an unpleasant feeling arousing while a person is not able to perform a task or cannot satisfy their need), *short term excitement* (experienced when the subject feels the psychological arousal of positive value), *meditation* (represents a person's composure and calmness), *engagement* (experienced when the subject is alert and consciously directs attention towards task-relevant stimuli) and *long term excitement* (reflects a person's general mood or emotional state, rather than reactions to short surprising stimuli) (Fig. 2).

Another device that can be used to gather emotions from the participants is thermal imaging camera. Emotion recognition using such device as compared to other methods is a very practical way, because it is non-invasive [8, 20, 22].

Fig. 2 Eyetracker example (on the *left side*) and an eyetracking heat map (on the *right side*)

Number of studies has been conducted, that have measured the impact of positive and negative stimulus on the temperatures of specific regions of the face [23, 24]. The conclusion of these studies was, among others, that negative emotions cause a lowering of the temperature of the nose (Fig. 3).

Other biometric devices planned to be used in our research are devices recording electrodermal activity (EDA) [17]. They record the electrical resistance of the skin, dependent on the degree of moisture-induced changes in the activity of the sweat glands, which is controlled by the sympathetic system. These changes are treated as a symptom of experiencing emotional response to the stimuli. It is also planned to use devices for measuring and recording the person's pulse.

In addition, our system will also collect messages that user enters into the system. This information will be analysed for the emotional value of the words that have been used to formulate them. This analysis will be carried out using Słowosieć [18] which is lexical-semantic base for Polish language. It contains a collection of synonyms (synonymous lexical units) described by short definitions expressed in

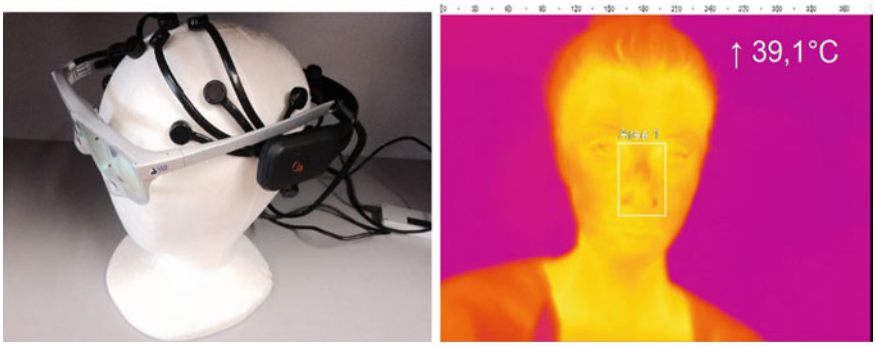


Fig. 3 EEG device example—Emotiv Epoc (on the *left side*) and the thermal image example with area for measuring the temperature set around the nose of the participant (on the *right side*)

natural language. Słowosieć serves as a dictionary where each individual concepts and meanings of words (lexical units) are defined by the place in the network of mutual relations, reflecting the lexical system of Polish language. Some of these concepts has also been described by the names of emotions that usually accompany the use of such terms and one of the fifth expression estimate its polarity like: strong negative, weak negative, neutral, weak positive, strong positive. We assign for each expression the score: *−2* points for strong negative, *−1* point for weak negative, *0* for neutral, *1* point for weak positive and *2* points for strong positive.

Finally, the user's emotional state is expressed by the sum of the points given for each word put in the system divided by the number of words in a message. Then students' emotional state has the value from [−2, 2] and it is interpreted in the following way: numbers near −2 mean that user has a very negative attitude, numbers oscillating around 0 mean that user is characterized by the neutral emotional state, numbers trending to 2 characterize persons with positive emotions.

In this paper we assume that a learner profile is represented as a tuple of values defined in the following way:

$$t: A \rightarrow V \tag{1}$$

where: A is a finite set of the profile attributes and V—the domain of all attributes from

$$A,\ V = \bigcup_{a \in A} V_a,\ \underset{a \in A}{\forall} (t(a) \in V_a). \tag{2}$$

The table shown below contains the content of learner profile and the typical methods for acquiring this type of data [14] (Table 1).

4 Summary

In order to increase the efficiency of the education process, the personalization and adaptation of intelligent tutoring courses should be applied. Our previous studies show [13] that fitting the educational material to the needs and preferences of the user increases their motivation to learn, and thus students achieve better learning outcomes in a shorter time. Personalization of the learning process should be provided at every stage of the education process. In the first, most important step, a system for remote learning should gather information about the users and identify their needs, preferences, learning styles, interests and personality traits.

The scope of the collected data will enable to strongly personalize [15] the process of recommendation of teaching materials for the user. We will use not only biometric data (such as eye movements, heart rate, blood pressure), but also their correlation with emotional state of the student. The latter will be estimated based on his facial expressions and the language he used during the interaction with the

Table 1 Content of learner profile

Type of data	Data	Form of acquiring data
Demographic data	Login, password, first name, second name, address, e-mail, telephone, age, sex, IQ, educational level	Questionnaire
Learning style (ILS)	Perception, receiving, processing, understanding	ILS Questionnaire
Abilities (IA)	Verbal comprehension, word fluency, computational ability, spatial visualization, associative memory, perceptual speed, reasoning	Psychological tests
Personal Character Traits (PCT)	Concentration, motivation, ambition, self-esteem, level of anxiety, locus of control, open mind, impetuosity, perfectionism, independence	Psychological tests
Interests (I)	Humanistic science, formal science, the natural science economics and law, technical science, business and administration, sport and tourism, artistic science, management and organization, education	Psychological tests
Biometrical Data (BD)	Time of eyes' focus on the specific content, overload, ignored learning material, most interesting learning material	Eyetracker device
	Frustration, short term excitement, meditation, engagement, long term excitement	EEG device—Emotiv Epoc
	Negative emotions	Thermal imagine camera
	Experiencing	Recording electrodermal activity and pulse
	General emotional state	Słowosieć
Usage data	Current and finished learning scenario time of learning with the reference to each learning material, test's score, the number of failed tests	Observations and saving of student's interaction with the system

system. This data can have a big impact on the results obtained in the course of teaching and its efficiency defined as the time required to master a certain material. The collected information about students are further used to learning scenario determination and modification.

Our designed system can be used in educational institutions (schools, universities) as an alternative to traditional classes in classrooms to conduct remedial classes for weak students for outstanding students as a form of additional development of their skills and expanding knowledge, which fails to convey in the traditional classes.

In addition, such a system can be very useful to train employees for any company, which saves time (employees can improve their education at a time that suits

them) and finance (once prepared, such training can be used for many years for many employees and without any help and supervision of additional people).

Our future works will be devoted to working out method of determination and modification learning scenario based on all collected data. Further, the proposed model of intelligent tutoring system will be implemented and verified by the real users.

References

1. Bajrakrarevic, N., Hall, W., Fullick, P.: Incorporating learning styles in hypermedia environment: empirical evaluation (2003). http://wwwis.win.tue.nl/ah2003/proceedings/paper4.pdf. Accessed 13.05.2016
2. Calviv, C., Porta, M., Sacchi, D.: e5Learning, an e-learning environment based on eye tracking. In: Proceedings of the 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008), pp. 376–380. Santander, Spain. Accessed 1–5 July 2008,
3. Chen, W., Mizoguchi, R.:, Leaner model ontology and leaner model agent. In: Kommers, P. (ed.) Cognitive Support for Learning—Imagining the Unknown, pp. 189–200. IOS Press (2004)
4. Chynal, P., Sobecki, J., Szymański, J.M.: Remote usability evaluation using eye tracking enhanced with intelligent data analysis. In: Design, User Experience, and Usability. Design Philosophy, Methods, and Tools, pp. 212–221. Springer, Berlin, Heidelberg (2013)
5. Emotiv EPOC User Manual. Emotiv (2009)
6. Graff, M.: Cognitive style and hypertext structures. In: Proceedings of the 4th European Learning Styles Conference (1999)
7. Grigoriadou, M., Papanikolaou, K., Kornilakis, H., Magoulas, G.: (2001), INSPIRE: An INtelligent System for Personalized Instruction in a Remote Environment. In: Lecture Notes In Computer Science, vol. 2266, pp. 215–225 (2001)
8. Gunes, H.: Automatic, dimensional and continuous emotion recognition (2010)
9. Gütl, C., Pivec, M., Trummer, C., García-Barrios, V.M., Mödritscher, F., Pripfl, J., Umgeher, M.: AdELE: a framework for Adaptive E-Learning through Eye Tracking. In: Proceedings of IKNOW (2004)
10. Harrison, T.: The Emotiv mind: investigating the accuracy of the Emotiv EPOC in identifying emotions and its use in an Intelligent Tutoring System (2013)
11. ILS Questionnaire. https://www.engr.ncsu.edu/learningstyles/ilsweb.html. Accessed 13 May 2016
12. Kobsa, A., Koenemann, J., Wolfgang, P.: Personalized hypermedia presentation techniques for improving online customer relationships. Knowl. Eng. Rev. **16**(2), 111–155 (2001)
13. Kozierkiewicz-Hetmańska, A.: Effectiveness of intelligent tutoring system offering personalized learning scenario. In: LNCS, vol. 7196, pp. 310–319 (2012)
14. Kozierkiewicz, A.: Content and structure of learner profile in an intelligent E-learning system, In: Knowledge Processing and Reasoning for Information Society, EXIT Warsaw, pp. 101–116 (2008)
15. Kozierkiewicz-Hetmańska, A., Nguyen, N.T.: A framework for building intelligent tutoring systems, In: Nguyen, N.T., Van Do, T., Le Thi, H.A. (eds.): Advanced Computational Methods for Knowledge Engineering. Studies in Computational Intelligence, vol. 479, s. 251–265. Springer, Heidelberg [i in.] (2013). ISSN 1860-949X
16. Kukla, E., Nguyen, N.T., Daniłowicz, C., Sobecki, J., Lenar, M.: A model conception for optimal scenario determination in an intelligent learning system. In: ITSE—Int. J. Interact. Technol. Smart Educ. **1**(3), 171–184 (2004)

17. Martini, F., Bartholomew, E.: Essentials of Anatomy & Physiology, s. 267. Benjamin Cummings, San Francisco (2003)
18. Maziarz, M., Piasecki, M., Szpakowicz, S.: Approaching plWordNet 2.0. In: Proceedings of the 6th Global Wordnet Conference. Matsue, Japan. Accessed 9–13 Jan 2012
19. Pask, G.: Styles and strategies of learning. Br. J. Educ. Psychol. No. 46, str. 128–148 (1976)
20. Pavlidis, I., Levine, J.: Thermal image analysis for polygraph testing. Eng. Med. Biol. Mag. IEEE **21**(6), 56–64 (2002)
21. Porta, M., Ricotti, S., Perez, C.J.: Emotional e-learning through eye tracking. In: Global Engineering Education Conference (EDUCON), pp. 1–6. IEEE. doi:10.1109/EDUCON.2012.6201145. Accessed 17–20 Apr 2012
22. Puri, C., Olson, L., Pavlidis, I., Levine, J., Starren, J.: StressCam: non-contact measurement of users' emotional states through thermal imaging. In: CHI'05 Extended Abstracts on Human Factors in Computing Systems, pp. 1725–1728. ACM, Apr 2005
23. Rimm-Kaufman, S.E., Kagan, J.: The psychological significance of changes in skin temperature. Motiv. Emot. **20**(1), 63–78 (1996)
24. Salazar-López, E., Domínguez, E., Ramos, V.J., de la Fuente, J., Meins, A., Iborra, O., Gómez-Milán, E.: The mental and subjective skin: emotion, empathy, feelings and thermography. Conscious. Cogn. **34**, 149–162 (2015)
25. Schwartz, B.E.: The advantages of digital over analog recording techniques. Electroencephalogr. Clin. Neurophysiol. **106**(2), 113–1117 (1998)
26. Sikka, R., Dhankhar, A., Rana, C.: A survey paper on e-learning recommender system. Int. J. Comput. Appl. **47**(9), 27–30 (2012)
27. Tian, F., Zheng, Q., Gong, Z., Du, J., Li, R.: Personalized learning strategies in an intelligent e-learning environment. In: Proceedings of the 2007 11th International Conference on Computer Supported Cooperative Work in Design (2007)
28. Tuaksubun, Ch., Mungsing, S.: Design of an intelligent tutoring system that comprises individual learning and collaborative problem-solving modules. In: Special Issue of the International Journal of the Computer, the Internet and Management, vol. 15, No. SP3 (2007)
29. Venkatesh, R., Naganathan, E.R., Uma Maheswari, N.: Intelligent tutoring system using hybrid expert system with speech model in neural networks. Int. J. Comput. Theory Eng. **2**(1), 1793–8201 (2010)
30. Wald, M., Wills, G., Millard, D., Gilbert, L., Khoja, S., Kajaba, J., Li, Y., Singh P.: Enhancing Learning using Synchronised Multimedia Annotation, Eunis (2009)
31. Yeaxlee, B.A.: Lifelong Education. Cassell, London (1929)

Two-Step Reduction of GOSCL Based on Subsets Quality Measure and Stability Index

Peter Butka, Jozef Pócs and Jana Pócsová

Abstract Generalized One-Sided Concept Lattices (GOSCL) represent a tool for extraction of hidden hierarchical structure among the datasets with different types of attributes. The specific problem of this method is an interpretation of the results from large created hierarchies, what often leads to the selection of the most relevant concepts. Subsets quality measure and stability index are techniques used for the ranking of the concepts relevance. In this paper we describe an approach which combines these two ranking techniques. The proposed approach is illustrated by an example and the experiments with the effect of reduction on generated input data tables are also provided.

Keywords Generalized one-sided concept lattices · Formal concept analysis · Reduction methods · Data analysis

P. Butka (✉)
Faculty of Electrical Engineering and Informatics, Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Letná 9, 04200 Košice, Slovakia
e-mail: peter.butka@tuke.sk

J. Pócs
Faculty of Science, Department of Algebra and Geometry, Palacký University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: pocs@saske.sk

J. Pócs
Mathematical Institute, Slovak Academy of Sciences, Grešákova 6, 040 01 Košice, Slovakia

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

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_36

1 Introduction

One of the several areas containing the approaches for identification of conceptual models from the datasets is called Formal Concept Analysis (FCA, [10]). It is suitable for creation of concepts hierarchies (known as concept lattices) from input data tables. This method has been successfully applied in several areas like conceptual data analysis, information retrieval, or data/text mining. In its basic version, so-called crisp case, FCA processes input data tables in the form of object-attribute models represented by the binary relations (an object has/has-not an attribute). As FCA has been proved as a useful tool, several extensions were proposed in order to process object-attribute models, where relationship between objects and their attributes is based on the fuzzy relations, cf. [1, 17, 19].

For the purposes of practical data mining tasks, certain fuzzy FCA models known as one-sided concept lattices were introduced. In this case object clusters are considered as ordinary subsets (as in the crisp case), while attributes are assumed to obtain fuzzy values [13]. In order to work with input data tables with different types of attributes an approach called Generalized One-Sided Concept Lattice (GOSCL) was proposed [6]. This model produces a hierarchical structure of concepts in the form of a concept lattice, which can be used in several applications for analysis of input data, e.g., [8, 11, 12].

Although outputs of FCA based methods in the form of concept lattices provide valuable information about considered object-attribute models, in many cases the resulting hierarchical structure is too complex and some reduction is needed. Several approach were already introduced, cf. [2, 14]. Hence, in order to select the most relevant concepts, some types of post-processing methods are applied. One of such methods are the so-called ranking-based methods. For GOSCL two such methods were proposed, in particular the subsets quality measure [7] and the intent stability index [9].

In this paper we provide an idea for the combination of these methods, where we use the subsets quality measure as a primary tool for the reduction of a hierarchical structure of all concepts. Consequently, the second step is based on the ranking of concepts in this reduced model by the intent stability index in order to get concepts with more similar values of their attributes. In the following section we provide the information on generalized one-sided concept lattices as well as basics of the mentioned concept ranking methods. Section 3 is devoted to the proposed approach for combination of these ranking methods and reduction of the original GOSCL model. The last section contains an example of such reduction and description of simple experiments based on the generated input data tables.

2 GOSCL Models and Selected Ranking Measures

In this section we provide some details regarding the model of GOSCL, as well as basic definitions of relevance measures which will be combined in our approach for reduction of concept lattices—subsets quality measure and stability index of concepts.

2.1 Basics of Generalized One-Sided Concept Lattices

Generalized One-Sided Concept Lattices (GOSCL) introduced in [6] are convenient for analysis of object-attribute models with different types of attributes. Hence we have to formalize this types of object-attribute models first. This is done by the notion of formal context.

A 4-tuple $\left(B, A, \mathsf{L}, R\right)$ is said to be a *generalized one-sided formal context* if the following conditions are fulfilled:

(1) B is a non-empty set of objects and A is a non-empty set of attributes.
(2) $\mathsf{L} : A \to \mathsf{CL}$ is a mapping from the set of attributes A to the class of all complete lattices. Hence, for an attribute $a \in A$, $\mathsf{L}(a)$ denotes a complete lattice, which represents a structure of truth values associated to the attribute a.
(3) R is generalized incidence relation (input data table), i.e., $R(b, a) \in \mathsf{L}(a)$ for all $b \in B$ and $a \in A$. Thus, $R(b, a)$ represents a degree from the structure $\mathsf{L}(a)$ in which the object b has the attribute a.

The next step is to define the concept forming operators, which form a Galois connection between the classical subsets of the set of all objects $\mathbf{P}(B)$ and the direct products of complete lattices $\prod_{a\in A} \mathsf{L}(a)$ presented in a formal context. This direct product can be seen as a generalization of the universe of fuzzy subsets over the attribute set A. However, in this case each attribute of A attains truth values from the complete lattice $\mathsf{L}(a)$.

Given a generalized one-sided formal context $\left(B, A, \mathsf{L}, R\right)$, the concept forming operators $^{\perp} : \mathbf{P}(B) \to \prod_{a\in A} \mathsf{L}(a)$ and $^{\top} : \prod_{a\in A} \mathsf{L}(a) \to \mathbf{P}(B)$ are defined by

$$X^{\perp}(a) = \bigwedge_{b\in X} R(b, a), \tag{1}$$

$$g^{\top} = \{b \in B : \forall a \in A,\ g(a) \leq R(b, a)\}. \tag{2}$$

Formal concepts are defined as the fixed points of the concept forming operators. Formally, $\mathscr{C}\left(B, A, \mathsf{L}, R\right)$ is defined as the set of all pairs (X, g), where $X \subseteq B$, $g \in \prod_{a\in A} \mathsf{L}(a)$, satisfying $X^{\perp} = g$ and $g^{\top} = X$. Set X is usually referred as the *extent* and g as the *intent* of a concept (X, g). Moreover, we can define the partial order on $\mathscr{C}\left(B, A, \mathsf{L}, R\right)$ as:

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

Finally, if $\big(B, A, \mathsf{L}, R\big)$ is a generalized one-sided formal context, then the set of all concepts $\mathscr{C}\big(B, A, \mathsf{L}, R\big)$ with the partial order defined above forms a complete lattice, which is also called generalized one-sided concept lattice corresponding to the formal context (object-attribute model) $\big(B, A, \mathsf{L}, R\big)$.

2.2 Relevance of Concepts—Quality Measure on Subsets of Objects

The quality measure of subsets for generalized one-sided concept lattices, primary as a tool for ranking a significance of extracted concepts, was proposed in [7].

Recall that a mapping $h\colon P \to Q$ between the partially ordered sets is said to be order-preserving if $x \leq y$ in P implies $h(x) \leq h(y)$ in Q. An order preserving mapping $h\colon \mathsf{L}(a) \to \mathbf{2}$, $\mathbf{2}$ denoting the two-element lattice with $0 < 1$, is called an h-cut. Any h-cut divides each partially ordered set into two respective parts, the pre-images $h^{-1}(0)$ and $h^{-1}(1)$, the first one being downward closed and the second one being upward closed.

Let $\big(B, A, \mathsf{L}, R\big)$ be a formal context and $\mathscr{H}_m = \big((h_a^n)_{a\in A}\big)_{n=1}^m$ be a system of $m \cdot |A|$ h-cuts, where $h_a^m\colon \mathsf{L}(a) \to \mathbf{2}$ for all $n = 1, \ldots, m$ and for all $a \in A$. Note that many times in practice it is convenient to select a threshold value $t \in \mathsf{L}(a)$ and consequently h_a^n is defined with respect to this value, e.g., $h_a^n(x) = 1$ if $x \geq t$ and $h_a^n(x) = 0$ otherwise. The system $\mathscr{H}_m$ naturally defines m binary relations indexed by $n = 1, \ldots, m$ where for given $b \in B$ and $a \in A$ we have $(b, a) \in R^n$ if and only if $h_a^n(R(b, a)) = 1$.

$$Q_m(X) = \frac{\Big|\big\{n \in \{1, \ldots, m\} \,:\, X^{\prime^n\prime^n} = X\big\}\Big|}{m} \tag{3}$$

Note that $'^n$ denotes the derivation operators associated to the formal context (B, A, R^n), see [10]. In this case, the condition $X^{\prime^n\prime^n} = X$ is equivalent to that X is an extents in the classical concept lattice determined by (B, A, R^n), i.e., there is $Y \subseteq A$ such that (X, Y) is a concept. The quality measure of a subset of objects represents frequency of its presence in concept lattices created from the binary contexts defined by particular h-cuts. From this point of view more relevant subsets have higher frequency, i.e., the quality measure Q_m.

2.3 Ranking of Concepts Based on Stability Index

Finally, within this subsection we briefly recall the basic notions concerning intent stability index for GOSCL, introduced in [9].

Let (B, A, L, R) be a generalized one-sided formal context. Given a concept $C = (X, g)$ of the generalized one-sided concept lattice $\mathscr{C}\big(B, A, \mathsf{L}, R\big)$, the intent stability index of this concept is defined in the following way:

$$\mathsf{S}_{\mathsf{tab}}\big(X, g\big) := \frac{\big|\{Y \in \mathbf{P}(X) \,:\, Y^{\perp} = g\}\big|}{2^{|X|}}. \tag{4}$$

The extent of a concept (X, g) is a maximal collection of objects satisfying given threshold g. Formally, the stability index of (X, g) is expressed as the ratio of the number of all subsets $Y \subseteq X$ for which $Y^{\perp} = g$, i.e., the threshold determined by objects in Y is precisely the same as for the whole extent set X, to the number of all possible subsets of X. Thus it represents the relative frequency of the event that a subset of a concept determines the same threshold value as a whole concept. This enables probabilistic interpretation of the stability index, $\mathsf{S}_{\mathsf{tab}}\big(X, g\big)$ equals to the probability that $Y^{\perp} = g$ for a random subset $Y \subseteq X$ chosen randomly with uniform probability $p = 0.5^{|X|}$. Consequently, Monte Carlo simulation (see Algorithm 1), can be used for an estimation of the parameter $\mathsf{S}_{\mathsf{tab}}\big(X, g\big)$, cf. [9] for more details.

Algorithm 1 Monte Carlo simulation for intent stability approximation

Require: Concept (X, g), number of repetition n
Ensure: Intent stability approximation $\mathsf{S}_{\mathsf{tab}}\big(X, g\big)$
1: $S := 0$
2: **for** $i = 1$ to n **do**
3: chose a random subset $Y \subseteq X$ ▷ An element $x \in X$ belongs to Y with probability 0.5
4: **if** $Y^{\perp} = g$ **then** $S := S + 1$
5: **end if**
6: **end for**
7: $\mathsf{S}_{\mathsf{tab}}\big(X, g\big) := S/n$
8: **return** $\mathsf{S}_{\mathsf{tab}}\big(X, g\big)$ ▷ Output of the algorithm

The parameter n represents number of repetitions of the Monte Carlo simulation. With respect to a required accuracy of the estimation, it can be obtained from the so-called Hoeffding inequality

$$\mathsf{Pr}\left(\left|\frac{S}{n} - \mathsf{S}_{\mathsf{tab}}(X, g)\right| \geq \varepsilon\right) \leq 2\mathrm{e}^{-2n\varepsilon^2} < \delta.$$

Parameter $\delta > 0$ represents a threshold value for probability of an absolute error $\varepsilon > 0$, i.e., after $n > \frac{\ln 2 - \ln \delta}{2\varepsilon^2}$ repetitions of Algorithm 1, the value S/n fulfills that the exact value $\mathsf{S}_{\mathsf{tab}}(X, g)$ lies in the open interval $\big(S/n - \varepsilon, S/n + \varepsilon\big)$ with probability at least $1 - \delta$.

3 Proposed Approach for Two-Step Reduction Procedure on GOSCL

In this section we provide the information on application of subsets quality measure and intent stability index as two steps in the reduction procedure. The first approach is generally based on the selection of fuzzy sets as h-cuts and computation of the frequency of concepts for all used h-cuts in their particular binary formal contexts. This approach is able to provide hierarchical structure (simplified version of original concept lattice) as a result of the reduction (if we select some threshold on quality measure). We have looked in more details on the resulted structure and we found that hierarchical structure constructed from the concepts selected for threshold t_Q is a complete lattice in case when $t_Q = 1.0$, as it is proved in Proposition 1. In other cases, even if the result is not the complete lattice, the structure is still concept hierarchy with the hierarchically organized structure of selected concepts and can be visualized as a reduced concept hierarchy with the ordering relation based on the subsets inclusion. Therefore, for the users or data analysts, the method of interpretation of concepts (clusters) characteristics (attributes) and their relations (connections between concepts) is same as in original GOSCL concept lattice.

Proposition 1 *Let* $\left(B, A, \mathsf{L}, R, \mathscr{H}_m\right)$ *be a generalized one-sided formal context with a system of h-cuts. The family* $\mathscr{S} = \{(X, g) \in \mathscr{C}(B, A, \mathsf{L}, R) : Q_m(X) = 1\}$*, i.e., the family of all concepts with extents of the subsets quality measure* $Q_m(X) = 1$*, forms a complete meet subsemilattice of the concept lattice* (B, A, L, R)*.*

Proof Recall that a subset $S \subseteq L$ of a complete lattice L is a meet subsemilattice, provided it is closed under arbitrary infima, i.e., $Y \subseteq S$ implies $\bigwedge Y \in S$.

Let $\{(X_i, g_i) : i \in I\} \subseteq \mathscr{S}$ be an indexed family of concepts. We show

$$\bigwedge_{i\in I} (X_i, g_i) = \left(\bigcap_{i\in I} X_i, \left(\bigvee_{i\in I} g_i\right)^{\top\perp}\right) \in \mathscr{S}.$$

In particular, it is sufficient to show that the subsets quality measure $Q_m\left(\bigcap_{i\in I} X_i\right)$ associated to the system $\mathscr{H}_m$ equals to one. As for all $i \in I$

$$Q_m(X_i) = \frac{\left|\left\{n \in \{1, \ldots, m\} : \exists Y \subseteq A, (X, Y) \in \underline{\mathfrak{B}}(B, A, R^n)\right\}\right|}{m} = 1,$$

for each $n \in \{1, \ldots, m\}$ there is $Y_i^n \subseteq A$ such that pair (X_i, A_i^n) forms a concept in $\underline{\mathfrak{B}}(B, A, R^n)$. However, since concept lattices $\underline{\mathfrak{B}}(B, A, R^n)$ are closed under infima, it follows that $\left(\bigcap_{i\in I} X_i, \left(\bigcup_{i\in I} X_i\right)''\right)$, $''$ denoting the composition of the derived operators for the classical concept lattices (see [10]), is a concept in $\underline{\mathfrak{B}}(B, A, R^n)$ as well. Consequently, for the subsets quality measure we obtain $Q_m\left(\bigcap_{i\in I} X_i\right) = 1$, which yields $\left(\bigvee_{i\in I} g_i\right)^{\top\perp}\right) \in \mathscr{S}$.

The second approach is based on the definition of stability, which prefers the concepts for which their subsets are usually very similar (according to their attributes) after usage of closure operator. The stability index leads to the selection of the stable concepts, but do not provide hierarchically organized structure as in case of subsets quality measure. Therefore, our main idea is to combine both methods and provide two-step approach for the reduction of the original GOSCL concept lattices. It means that we use subsets quality measure in order to provide hierarchy-based reduction first and then select the subset of the hierarchy based on the stability index.

Two-step Reduction Procedure

- **Input**: $\mathscr{C}(B, A, \mathsf{L}, R)$, system of h-cuts, threshold t_Q for Q_m, threshold t_S for $\mathsf{S}_{\mathsf{tab}}$
- **Step 1**—selection of sub-hierarchy of concepts $SH(t_Q)$ using Q_m, where every concept c with $Q_m(c) \geq t_Q$ is included
- **Step 2**—selection of concepts from sub-hierarchy $SH(t_Q)$, where every concept c with $\mathsf{S}_{\mathsf{tab}} \geq t_S$ is included
- **Output**: selected sub-structure of concepts from $SH(t_Q)$ based on $\mathsf{S}_{\mathsf{tab}}$

4 Illustrative Example and Experiments

Now we provide an example of the presented approach and the results of experiment with the reduction on generated datasets. For the purpose of this paper we used simple procedure for generating the dataset based on the ordinal attributes a_i with the values from $0, 1, 2, 3, 4$. It is simple generator for generalized one-sided context with three inputs—number of objects, number of attributes and ratio of sparseness s. The last parameter is used for setup of relative number of zeros in generated input data tables, with $s = 0.9$ number of zeros is 90 % of all values in data table.

In our illustrative example number of objects was 50, number of attributes 5 and we have used sparseness level $s = 0.5$. Of course, for the application of subsets quality measure we also need to setup h-cuts. For this purpose we have used the definition of cuts based on four fuzzy sets, which can be described by these vectors of threshold values: $(3, 3, 2, 3, 1)$, $(1, 3, 3, 2, 3)$, $(3, 1, 3, 3, 1)$, $(3, 3, 1, 3, 1)$. This leads to the system of four h-cuts and we are able to compute Q_m ranking of concepts of original GOSCL. When we select some threshold value t_Q, the original concept lattice will be reduced to conceptual sub-structure based on the selection of concepts according to this threshold.

For the illustrative example, one of the generated context originally produced the concept lattice with 154 concepts. Such concept lattice (see Fig. 1) is not very suitable for data analyst for understanding the results and it is useful to apply reduction techniques. Here, we applied our two-step reduction procedure. For the first step we used previously mentioned setup of h-cuts to provide constraints on concepts and therefore we applied subsets quality measure for simplification of the conceptual structure. For the threshold $t_Q = 0.5$ we get conceptual model depicted in Fig. 2 with 20 concepts, which surpassed the constraints defined by user-defined cuts. The

Fig. 1 Example of original concept lattice on generated dataset for sparseness $s = 0.5$ with 50 objects and 5 chain-based attributes

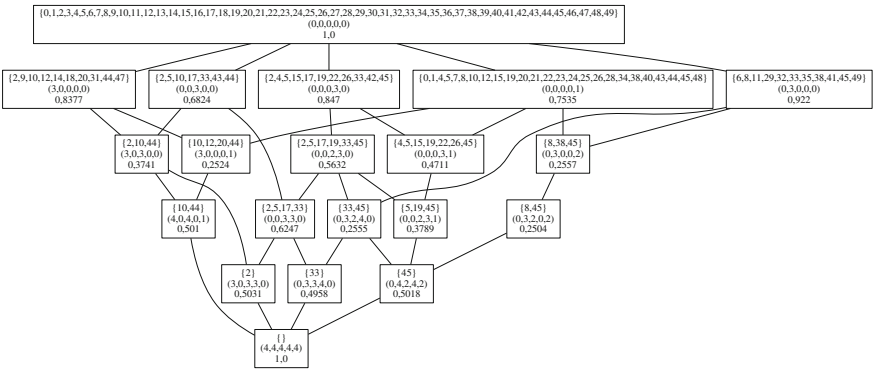


Fig. 2 Reduced conceptual sub-structure based on first step of our approach—subsets quality measure. The data were generated for 50 objects and 5 chain-based attributes, reduction was based on reduction threshold $t_Q = 0.5$ and four selected h-cuts (stability index values are added to concepts)

concepts in figure are represented by rectangles with first row containing the index of objects for concepts (extent) and second row containing the attributes values (intent). In order to better show the next step, we have also added stability index values as the third row in every concept.

While for this illustrative example the resulted conceptual structure has reasonable size, for larger original lattice or differently defined settings it can be still difficult to understand the data clusters and their connections. As concept stability idea proved to be useful for the selection of the most interesting concepts from the concept lattices, we decided to apply stability index threshold as a second step in our reduction procedure. In our illustrative example we selected concepts with better stability index than $t_S = 0.5$, the resulted structure is shown in Fig. 3. For the comparison you can see the difference to previous step thanks to already shown stability indexes of concepts in reduction after the first step. As we can see, final reduced model contains 12 concepts, which are both better to removed ones according to the selected thresholds in subsets quality measure and stability index.

Now we provide some results from the experiments we did based on the generated input data tables. In this case we have used fixed number of 10 attributes, sparseness $s = 0.5$ and also h-cuts were fixed to one specific set of cuts (in order to have comparable results for changing parameters). The watched value in all experiments was reduction ratio defined as number of concepts for particular reduction method

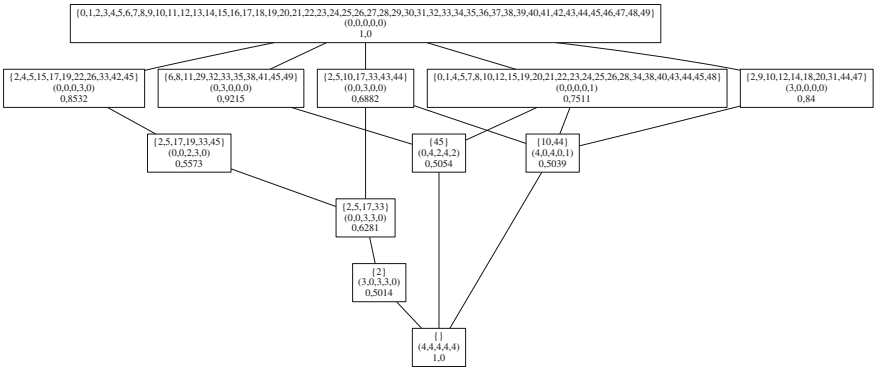


Fig. 3 Final conceptual structure generated from the original concept lattice after our two-step reduction procedure (with the intent stability index threshold $t_s = 0.5$)

divided by the number of concept in original GOSCL. The results are presented in Fig. 4. We have analyzed the reduction ratio of subsets quality measure (Quality in graphs), intent stability index (Stability), and our approach with the combination of both techniques (Quality + Stability). The experiments were done according to different number of objects (from 25 to 100) and for four different combinations of

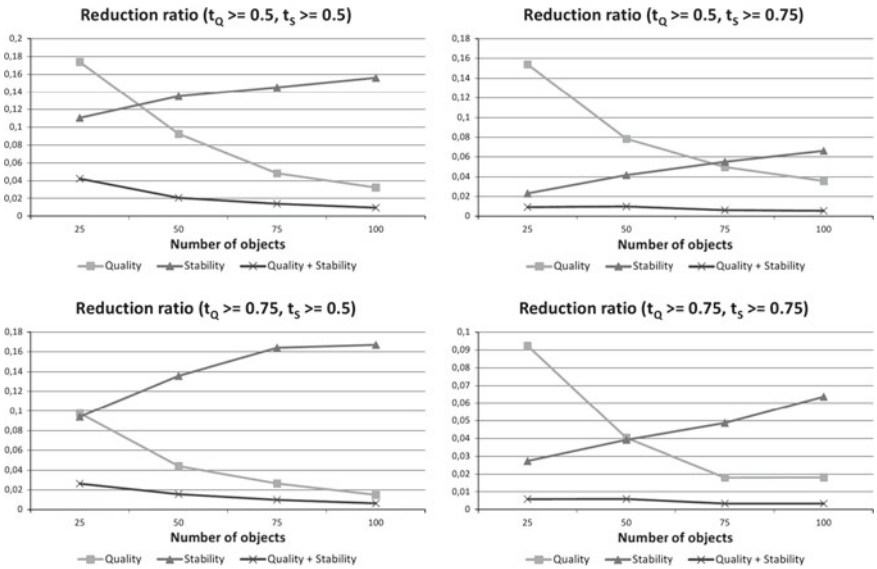


Fig. 4 The results of experiments—reduction ratio for subsets quality measure (Quality), intent stability index (Stability), and their combination, according to number of objects and four different combinations of thresholds t_Q and t_S (number of attributes, cuts setup and sparseness of the generated data tables were fixed)

thresholds t_Q and t_S (depicted in particular graphs within the figure). As we can see from the graphs, intent stability index reduction is smaller with the increased number of objects, but the reduction based on Q_m is stronger and therefore combined approach is able to reduce the growing concept lattice (with the number of objects) more. As the consequence of this fact, combined approach is able to provide quite small number of selected concepts in the final hierarchical structure.

In practical application of the proposed approach data analyst will be important in the identification of *h*-cuts setup for particular domain of the analysis. In the future we would like to test our approach as a selection method for relevant concepts on real datasets in different domains like clustering of textual documents [20, 21], information retrieval based on concept lattices [15], analysis of e-learning users groups [4, 18], meteorological data [5], or data about patients from the medical domain [3, 16]. We also want to study the statistical dependencies between both particular reduction techniques.

5 Conclusions

In this paper we have presented the combined approach for the reduction of concepts from generalized one-sided concept lattices by the usage of subsets quality measure and intent stability index. The proposed approach can be useful as a tool for the selection of relevant clusters from the analysis of input data tables. We have provided also some experiments in order to show the reduction ratios of the combined approach and particular methods on the generated data tables.

Acknowledgments The first author was supported by the Slovak VEGA Grants 1/0493/16 and by the Slovak KEGA grant 025TUKE-4/2015. The second author was supported by the Slovak VEGA Grant no. 2/0044/16 and by the IGA project of the faculty of Science Palacký University Olomouc no. PrF2015010. The third author was supported by the Slovak Research and Development Agency under the contracts No. APVV-14-0892, VEGA grants No. 1/0529/15, No. 1/0908/15 and KEGA grant 040TUKE-4/2014.

References

1. Antoni, L., Krajči, S., Krídlo, O., Macek, B., Pisková, L.: On heterogeneous formal contexts. Fuzzy Set. Syst. **234**, 22–33 (2014)
2. Antoni, L., Krajči, S., Krídlo, O.: Stability of extents in one-sided fuzzy concept lattices. In: Proceedings of ITAT 2015, CEUR Workshop Proceedings, vol. 1422, pp. 3–8 (2015)
3. Babic, F., Majnaric, L., Lukacova, A., Paralic, J., Holzinger, A.: On patients characteristics extraction for metabolic syndrome diagnosis: predictive modelling based on machine learning. Lect. Notes Comput. Sci. **8649**, 118–132 (2014)
4. Babic, F., Paralic, J., Bednar, P., Racek, M.: Analytical framework for mirroring and reflection of user activities in E-learning environment. Adv. Intell. Soft Comput. **80**, 287–296 (2010)
5. Bartok, J., Babic, F., Bednar, P., Paralic, J., Kovac, J., Bartokova, I., Hluchy, L., Gera, M.: Data mining for fog prediction and low clouds detection. Comput. Inf. **31**(6+), 1441–1464 (2012)

6. Butka, P., Pócs, J.: Generalization of one-sided concept lattices. Comput. Inf. **32**(2), 355–370 (2013)
7. Butka, P., Pócs, J., Pócsová, J.: Reduction of concepts from generalized one-sided concept lattice based on subsets quality measure. Adv. Intell. Syst. Comput. **314**, 101–111 (2015)
8. Butka, P., Pócs, J., Pócsová, J.: Distributed computation of generalized one-sided concept lattices on sparse data tables. Comput. Inf. **34**(1), 77–98 (2015)
9. Butka, P., Pócs, J., Pócsová, J.: On intent stability index for one-sided concept lattices. In: Proceedings of 10th Jubilee IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI 2015), pp. 79–84 (2015)
10. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin (1999)
11. Halaš, R., Pócs, J.: Generalized one-sided concept lattices with attribute preferences. Inf. Sci. **303**, 50–60 (2015)
12. Kardoš, F., Pócs, J., Pócsová, J.: On concept reduction based on some graph properties. Knowl. Based Syst. **93**, 67–74 (2016)
13. Krajči, S.: Cluster based efficient generation of fuzzy concepts. Neural Netw. World **13**(5), 521–530 (2003)
14. Kumar, ChA, Srinivas, S.: Concept lattice reduction using fuzzy K-Means clustering. Expert Syst. Appl. **37**(3), 2696–2704 (2010)
15. Kumar, C.A., Mouliswaran, S.C., Amriteya, P., Arun, S.R.: Fuzzy formal concept analysis approach for information retrieval. Adv. Intell. Syst. Comput. **415**, 255–271 (2015)
16. Lukacova, A., Babic, F., Paralicova, Z., Paralic, J.: How to increase the effectiveness of the hepatitis diagnostics by means of appropriate machine learning methods. Lect. Notes Comput. Sci. **9267**, 81–94 (2015)
17. Medina, J., Ojeda-Aciego, M., Ruiz-Calviño, J.: Formal concept analysis via multi-adjoint concept lattices. Fuzzy Set. Syst. **160**, 130–144 (2009)
18. Paralič, J., Richter, C., Babič, F., Wagner, J., Raček, M.: Mirroring of knowledge practices based on user-defined patterns. J. Univ. Comput. Sci. **17**(10), 1474–1491 (2011)
19. Pócs, J., Pócsová, J.: Basic theorem as representation of heterogeneous concept lattices. Front. Comput. Sci-Chi. **9**(4), 636–642 (2015)
20. Sarnovský, M., Ulbrik, Z.: Cloud-based clustering of text documents using the GHSOM algorithm on the GridGain platform. In: Proceedings of 8th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI 2013), pp. 309–313 (2013)
21. Sarnovský, M., Čarnoká, N.: Distributed algorithm for text documents clustering based on k-Means approach. Adv. Intell. Syst. Comput. **430**, 165–174 (2016)

Author Index

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2