Dmitry I. Ignatov, Mikhail Yu. Khachay, Alexander Panchenko, Natalia Konstantinova, Rostislav Yavorsky, Dmitry Ustalov (Eds.)

AIST'2014 — Analysis of Images, Social Networks and Texts

Supplementary Proceedings of the 3rd International Conference on Analysis of Images, Social Networks and Texts (AIST'2014)
April 2014, Yekaterinburg, Russia

The proceedings are published online on the CEUR-Workshop web site in a series with ISSN 1613-0073, Vol-1197.

Preface

This volume contains supplementary proceedings of the third conference on Analysis of Images, Social Networks, and Texts (AIST'2014). The first two conferences in 2012 and 2013 attracted a significant number of students, researchers, academics and engineers working on data analysis of images, texts and social networks.

The broad scope of AIST makes it an event where researchers from different domains, such as image and text processing, exploiting various data analysis techniques, can meet and exchange ideas. We strongly believe that this may lead to cross-fertilisation of ideas between researchers relying on modern data analysis machinery. Therefore, AIST brings together all kinds of applications of data mining and machine learning techniques. The conference allows specialists from different fields meet each other, present their work and discuss both theoretical and practical aspects of their data analysis problems. Another important aim of the conference is to stimulate scientists and people from the industry to benefit from the knowledge exchange and identify possible grounds for fruitful collaboration.

Following an already established tradition, the conference was held in Yekaterinburg, capital of Urals region in Russia on 10-12 April 2014. The key topics of AIST are analysis of images and videos; natural language processing and computational linguistics; social network analysis; machine learning and data mining; recommender systems and collaborative technologies; semantic web, ontologies and their applications; analysis of socio-economical data.

The Programme Committee and reviewers of the conference featured well-known experts in data mining, natural language processing, image processing and related areas from Russia and all over the world.

We have received 74 high quality submissions mostly from Russia and also from France, Germany, India, Poland, Spain, Ukraine and USA, among which only 21 papers have been accepted as regular oral papers (11 long and 10 short) in Communications in Computer and Information Science series (Springer), volume 436. Thus, the acceptance rate of this volume is roughly 28%. In order to stimulate young practitioners and researchers we have also included 3 short industry papers in the main volume and 33 submissions as posters in this supplementary proceedings. Each submission has been reviewed by at least 3 reviewers, who are experts in their field, in order to supply detailed and helpful comments.

The conference also featured several invited talks and tutorials, as well as an industry session describing current trends and challenges.

Invited talks:

- Boris Mirkin (Higher School of Economics, Moscow), Data Clustering: Some Topics of Current Interest
- Jaume Baixeries (Universitat Politècnica de Catalunya, Barcelona), Characterization of Database Dependencies with FCA and Pattern Structures
- Dmitriy Kolesov (NextGIS, Moscow), GIS as an Environment for Integration and Analysis of Spatial Data
- Natalia Konstantinova (University of Wolverhampton, UK), Relation Extraction – Let's Find More Knowledge Automatically

Tutorials:

- Jaume Baixeries (Universitat Politècnica de Catalunya, Barcelona), Introduction to Formal Concept Analysis and Attribute Dependencies (in 2 parts)
- Konstantin Voronstov (CCAS of RAS and Yandex, Moscow), Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization
- Natalia Konstantinova (University of Wolverhampton, UK), Introduction to Dialogue Systems, Personal Assistants are Becoming a Reality
- Natalia Konstantinova (University of Wolverhampton, UK), Academic Writing – Getting Published in International Journals and Conferences

The industry speakers covered rich variety of topics:

- Iosif Itkin (Exactpro Systems), Network Models for Exchange Trade Analysis
- Leonid I. Levkovitch-Maslyuk (EMC), Big Data for business
- Alexander Semenov (http://jarens.ru/), Recent Advances in Social Network Analysis
- Irina Radchenko (http://iRadche.ru), Current Trends in Open Data and Data Journalism
- Oleg Lavrov (KM Alliance Russia), Knowledge Management as a Link between Business & IT
- Yury Kupriyanov (WikiVote!), IT Trends and Challenges in Knowledge Management

We would like to mention the best conference paper, according to the PC decision, written by V.B. Surya Prasath and Radhakrishnan Delhibabu and entitled "Automatic Contrast Parameter Estimation in Anisotropic Diffusion for Image Restoration".

We would like to thank the authors for submitting their papers and the members of the Programme Committee for their efforts to provide exhaustive reviews. We would also like to express special gratitude to all the invited speakers and industry representatives. We deeply thank all the partners and sponsors, and owe our gratitude to the Scientific Fund of Higher School of Economics for providing five AIST's participants with travel grants. Our special thanks goes to Springer editors who helped us starting from the first conference call until the final version of the proceedings. Last but not least, we are grateful to all organisers, especially to Eugeniya Vlasova and Dmitry Ustalov, whose endless energy saved us in the most critical stages of the conference preparation.

In Russian AIST is even more than abbreviation, namely it means a stork. So we believe this young and rapidly growing conference still be bringing inspiration for data scientists.

May, 2014

Dmitry I. Ignatov
Mikhail Yu. Khachay
Alexander Panchenko
Natalia Konstantinova
Rostislav Yavorsky
Dmitry Ustalov

Organisation

The conference was organized by joint team from the National Research University Higher School of Economics (HSE), Russia and Krasovsky Institute of Mathematics and Mechanics of UB RAS, Russia. It was supported by the Russian Foundation for Basic Research grant no. 14-07-06803.

Program Committee Chairs

Dmitry I. Ignatov	National Research University Higher School of Economics, Russia
Mikhail Yu. Khachay	Krasovsky Institute of Mathematics and Mechanics of UB RAS, Russia
Alexander Panchenko	Université catholique de Louvain, Belgium & Digital Society Laboratory, Russia

Organising Chair

Rostislav Yavorsky	National Research University Higher School of Economics, Russia

Proceedings Chair

Natalia Konstantinova	University of Wolverhampton, UK

Poster Chair

Dmitry Ustalov	Krasovsky Institute of Mathematics and Mechanics of UB RAS & Ural Federal University, Russia

Organising Committe

Eugeniya Vlasova	Higher School of Economics, Moscow
Dmitry Ustalov	Krasovsky Institute of Mathematics and Mechanics of UB RAS & Ural Federal University, Yekaterinburg
Olga Fazylova	Higher School of Economics, Moscow
Liliya Galimziyanova	Ural Federal University, Yekaterinburg
Alexandra Kaminskaya	Yandex, Moscow
Maria Khakimova	Perm State University, Perm
Natalia Korepanova	Higher School of Economics, Moscow
Elena Nikulina	Higher School of Economics, Moscow
Andrey Shelomentsev	APIO, Yekaterinburg and Moscow
Anna Voronova	Yandex, Moscow

Program Committee

Tinku Acharya	Videonetics Technology Pvt. Ltd.
Simon Andrews	Sheffield Hallam University, UK
Olga Barinova	Moscow State University & Yandex, Russia
Malay Bhattacharyya	University of Kalyani, India
Vladimir Bobrikov	Imhonet Research, Russia
Elena Bolshakova	Moscow State University & National Research University Higher School of Economics, Russia
Jean-Leon Bouraoui	Prometil SARL, France
Pavel Braslavski	Ural Federal University & Kontur Labs, Russia
Ekaterina Chernyak	National Research University Higher School of Economics, Russia
Ilia Chetviorkin	Moscow State University, Russia
Marina Chicheva	Image Processing Systems Institute of RAS, Russia
Florent Domenach	University of Nicosia, Cyprus
Alexey Drutsa	Moscow State University & Yandex, Russia
Maxim Dubinin	NextGIS, Russia
Victor Erukhimov	Itseez, Russia
Rashid Faizullin	Omsk State Technical University, Russia
Thomas Francois	Université catholique de Louvain, Belgium
Boris Galitsky	Become.com, USA
Dmitry Ilvovsky	Fors & Higher School of Economics, Russia
Vladimir Ivanov	Kazan Federal University, Russia
Vadim Kantorov	INRIA, France
Mehdi Kaytoue	LIRIS - INSA de Lyon, France
Vladimir Khoroshevsky	Dorodnicyn Computing Centre of RAS and National Research University Higher School of Economics, Russia
Sergei Koltsov	National Research University Higher School of Economics, Russia
Olessia Koltsova	National Research University Higher School of Economics, Russia
Natalia Konstantinova	University of Wolverhampton, UK
Anton Konushin	Lomonosov Moscow State University, Russia
Yuri Kudryavcev	PMSquare, Australia
Sergei Kuznetsov	National Research University Higher School of Economics, Russia
Victor Lempitsky	Skolkovo Institute of Science and Technology, Russia
Alexander Lepskiy	National Research University Higher School of Economics, Russia
Natalia Loukachevitch	Research Computing Center of Moscow State University, Russia
Julian Mcauley	Stanford University, USA

Vlado Menkovski	Eindhoven University of Tehnology, The Netherlands
Olga Mitrofanova	St. Petersburg State University, Russia
Dmitry Mouromtsev	National Research University of Information Technologies, Mechanics and Optics, Russia
Xenia Naidenova	Kirov Military Medical Academy, Russia
Sergey Nikolenko	Steklov Mathematical Institute of RAS & National Research University Higher School of Economics, Russia
Sergey Objedkov	Digital Society Lab & National Research University Higher School of Economics, Russia
Evgeniy Perevodchikov	Tomsk State University of Control Systems and Radioelectronics, Russia
Andrey Philippovich	Bauman Moscow State Technical University, Russia
Jonas Poelmans	Katholieke Universiteit Leuven, Belgium
Simon Polovina	Sheffield Hallam University, UK
Alexander Porshnev	National Research University Higher School of Economics, Russia
Irina Radchenko	National Research University Higher School of Economics, Russia
Delhibabu Radhakrishnan	RWTH Aachen, Germany
Konstantin Rudakov	Dorodnicyn Computing Centre of RAS, Russia
Andrey Savchenko	National Research University Higher School of Economics, Russia
Alexandra Savelieva	Microsoft & National Research University Higher School of Economics, Russia
Yuri Smetanin	Russian Foundation for Basic Research, Russia
Leonid Sokolinsky	South Ural State University, Russia
Rustam Tagiew	Qlaym Gmbh, Germany
Irina Temnikova	University of Wolverhampton, UK
Alexander Ulanov	HP Labs, Russia
Dmitry Vinogradov	All-Russian Institute for Scientific and Technical Information of RAS, Russia
Alexander Vodyaho	Saint Petersburg Electrotechnical University "LETI", Russia
Konstantin Vorontsov	Dorodnicyn Computing Centre of RAS, Russia
Leonind Zhukov	Ancestry.com, USA & National Research University Higher School of Economics, Russia
Nataly Zhukova	St.Petersburg institute for Infromatics and Automation of RAS, Russia
Dominik Ślęzak	University of Warsaw, Poland & Infobright Inc.

Invited Reviewers

Aleksei Buzmakov	INRIA, France

Yuri Katkov	Blue Brain Project, Switzerland
Lidia Pivovarova	University of Helsinki, Finland
Alexander Semenov	Higher School of Economics Alumni, Russia
Nikita Spirin	University of Illinois at Urbana-Champaign, USA
Dmitry Ustalov	Krasovsky Institute of Mathematics and Mechanics of UB RAS & Ural Federal University, Russia
Andrey Bronevich	Higher School of Economics Alumni, Russia
Sujoy Chatterjee	University of Kalyani, India
Surya Prasath	University of Missouri-Columbia, USA
Kirill Shileev	Digital Society Laboratory, Russia
Natalia Vassilieva	HP Labs, Russia
Victoria Yaneva	University of Wolverhampton, UK

Partners and Sponsoring Institutions

Central & Eastern European Software Engineering Conference in Russia
CLAIM
Data Mining Labs
Digital Society Laboratory
EMC
GraphiCon
Innovative Trading Systems
Krasovsky Institute of Mathematics and Mechanics
NextGIS
NLPub
penxy
Russian Foundation for Basic Research
SKB Kontur
Ural Federal University
Ural IT Cluster
WikiVote
Yandex

Table of Contents

Posters and Short Papers in English

Terminology Extraction from the Baidu Encyclopedia

Bulat Fatkulin[1,2]

[1] South Ural State University, Chelyabinsk, Russia
[2] Chelyabinsk State University, Chelyabinsk, Russia
bfatkulin@gmail.com

Abstract. The article examines the use of the applied linguistics technologies in the teaching of orientalistics in the Russian Federation higher education system. The research discloses the methods of the terminological units extracting using texts in Chinese studies of the modern Afghanistan. The author shows the solutions for intensive summarization and annotation of Chinese texts and language teaching methods for students to work with assistive software. The achievements of the Stanford NLP group are used for the Chinese text segmentation and named entity recognition.

Keywords: terminology extraction, orientalistics, natural language processing.

The applied linguistics can not be a "thing-in-itself", it serves determined interests. The teaching of Oriental and Asian languages occupies the special place in the system of Russian higher education. As a rule, the orientalistics education is obtained by students in elder age, being combined with certain tasks. Elitism and interdisciplinarity are key features of Orientalistics.

Variety of oriental cultures surrounding Russia is reflected in a wide range of Orientalistic branches (Iranian studies, Arabic studies, Turkology, Indology, Afgan studies etc.) The Sinology occupies a leading position among them [1]. All major world civilization centers, including Russia and China, have their own versions of orientalistics branches and use their own terminology. The following reasons make Afghan studies in China actual:

1. In recent years China has been active in developing countries in Asia and Africa and is a main investor in Afghanistan.
2. Afghan Studies became of great significance because of the region strategic location, since China relies heavily on the oil resources of the Middle East. The strategy of the Silk Road revival require control of eastern transport corridors.
3. The knowledge of the Islamic world terminology (including Afghan culture) is necessary to struggle against the religious extremism and the Uighur separatism.

China has its own political doctrine, information sources and media, the Chinese Internet is governed by the political doctrine of the country and provides information in accordance with its interests. Typological structure of the Chinese language and hyeroglyphs makes the direct borrowing of political terminology of the European languages impossible, and this feature deprives the external actors their possibility for public opinion manipulation. The Soviet heritage Middle Asian republics, Afghanistan, Pakistan and Iran are the closest western neighbors of China [2]. Due to many internal and external factors (ongoing civil war, the presence of the foreign military forces, drug trafficking, etc.) Afghanistan is a subject of attention to both Russia and China.

Russia should be aware it's Chinese ally projects, and therefore the study of the peculiarities of the Chinese terminology of Afghan studies can enrich Russian analytical networks with valuable experience. Studying of the Chinese Afghanistics terminology is necessary [9] for professionals involved in the work of intergovernmental organizations such as the SCO, BRICS, Custom Union, etc.

It should be stressed that Chinese experts use in their Afghan studies their own authentic terminology which largely differes from the terminology of the English-speaking global network structures and it's equivalents in European and Russian languages [8]. The philosophy of Confucianism and theoretical and methodological approaches of the Communist Party of China form the base of Chinese political terminology and are unknown for the wide range of scientists who do not understand Chinese.

If we intend to collect the relevant information we have to handle in a short time a large number of texts in the original language. Bare translation of Orientalistics articles from Western magazines impoverishes informational awareness. Qualified orientalists should be able to work independently with Chinese Oriental sources and must apply innovative educational technologies [4]. Fast annotation and summarization of texts is required from students. The access to these technologies develops the creative potential of them. Russian Terminography should work at the intersection of Chinese Studies [6], Islamic Studies,Arabic Studies and Iranian Studies. The high complexity of the Chinese texts processing necessitates the use of innovative technologies, which are based on the latest achievements of applied linguistics.

There are numerous methods of terminology extraction from large amounts of text, called corpora. The variety of algorithms and programs in different programming languages are used to exfor the term extraction. There are software products also ready-to-use for ordinary researchers. We have found a lot of information about algorythms of applied linguistic programs in articles of a Tomsk famous explorer O. S. Jacko [10].

Applied Linguistics for Chinese includes a wide range of specialized programs such as:

1. segmenters,
2. morphological analyzers,
3. parsers,
4. converters of encodings,

5. characters OCR systems,
6. databanks [3].

The text segmentation is automatically produced by the segmenter — a special program or script. Character is determined by segmenter task to get some information from the text analysis parameters are set in advance. The joined information is provided in a certain manner and conducted in one of the programming languages. Three phases are logical segmenting process stages: first it is punctual collection of information, for example, it may be a code web pages. Then, it is data analysis, processing and transformation into the desired format. Finally — it is providing result output.

In our work we used tools such as:

1. Stanford Chinese segmenter `http://nlp.stanford.edu:8080/parser/`
2. Shanghai Chinese language segmenter `http://hlt030.cse.ust.hk/research/c-assert/`
3. Automatic annotation of Chinese texts `http://www.chinese-tools.com`

It is difficult to overestimate the advantages of parser using for fast processing of the Chinese text are. The segmenter makes grouping of characters into combinations. The essence of this phenomenon can be explained by comparing the presentation of texts in Russian and Chinese. In Russian, the words are separated by spaces. However terminological combinations usually consist of several words. The stable combinations of words are easily recognized by native Russian languagers, but such grouping of words is difficulty for foreigners. In Chinese texts similar gaps stay between the standard characters. But the Chinese word can consist of multiple characters [5]. Terms, in turn, may consist of several words. Segmenter solves the problem of putting a space between characters groups, allows you to find the terms of several groups of characters.

To carry out the above-mentioned routine operations related to the recovery terminology, we used the Stanford Chinese segmenter, which uses probabilistic algorithms. The program is designed by Pi-Chuan Chang, Huihsin Tseng and Galen Andrew. we downloaded and installed this segmenter on a personal computer running operating system Linux Ubuntu. It works in Java 6 (JDK1.6)

Two segmentation models are provided. The "ctb" model was trained with Chinese treebank (CTB) segmentation, and the "pku" model was trained with Beijing University's (PKU) segmentation. PKU models provide smaller vocabulary sizes and OOV rates on test data than CTB models.

For both CTB and PKU, we provide two models representing slightly different feature sets. Models "ctb" and "pku" incorporate lexicon features to increase consistency in segmentation. The details of the segmenter can be found in the paper [12]. The description of the lexicon features can be found in [13].

The program runs from the command line by means of this command:

```
segment.sh [-k] [ctb | pku] <filename> <encoding> <size>
 ctb: Chinese Treebank
 pku: Beijing Univ.
```

The main principle of the Stanford segmenter is described in the work of Levy and Manning [11].

The Chinese text before the processing looked as follows:

比尔兼德高地，北部有厄尔布兹山脉，德马万德峰海拔5670米，为伊朗最高峰。西部和 西南部是宽阔的扎格罗斯山山系，约占国土面积一半。中部为干燥的盆地，形成许多沙 漠，有卡维尔荒漠与卢特荒漠，平均海拔1，000余米。仅西南部波斯湾沿岸与北部里海 沿岸有小面积的冲击平原。西南部扎格罗斯山麓至波斯湾头的平原称胡齐斯坦。

The same Chinese text after the processing segmenting has become much more clear:

尔 兼德 高地 ， 北部 有 厄尔布兹 山脉 ， 德马万德峰 海拔 5670 米 ， 为 伊朗 最高 峰 。 西部 和 西南部 是 宽阔 的 扎 格罗斯 山山 系 ， 约占 国土 面积 一半 。 中部 为 干燥 的 盆地 ， 形成 许多 沙漠 ， 有 卡维尔 荒漠 与 卢特 荒漠 ， 平均 海拔 1，000余 米 。 仅 西南部 波斯湾 沿岸 与 北部 里海 沿岸 有 小 面积 的 冲击 平原 。 西南部 扎 格罗斯 山麓 至 波斯湾 头 的 平原 称 胡齐斯坦 。

As we can see, the boundaries of Chinese words, consisting of several characters, are clearly marked.

At the second stage, using the method of regular expressions, we pulled the group of the received characters in a vertical chain, and then translated it with an automatic translator. In the third stage, we chose a combination, satisfying the requirements of the terms. Particular attention was paid to extract terms from titles chapters and subchapters section “Afghanistan” representing the ontology information. As a result the terms were broken up into meaningful groups to compile a thesaurus of the Chinese Afghan Studies.

The section “Afghanistan” of the Chinese online encyclopedia Baidu were chosen by us as the object of investigation. Baidu is online encyclopedia in Chinese, which develops and supports the Chinese search engine Baidu. As well as Baidu itself, the encyclopedia is censored in accordance with Chinese government regulations. On June 2013 Baidu encyclopedia contained more than 6.2 million articles (more than English and German Wikipedia together) and had more than 3.2 million of participants.

Our work was divided into several stages:

1. selection of raw texts about Afghanistan in Chinese,
2. using the word processing program for automatic annotation of the text and isolation of terminological phrases,
3. updating the terminology.

The ontology, the geographical names of Afghanistan in Chinese transcription, ethnonyms peoples of Afghanistan and Central Asia [6], the names of political figures of Afghanistan in the Chinese transcription, the terms of political geography, the names of international organizations [7], Islamic concepts in Chinese, Arabisms and Farsisms in Chinese transcription became the object of special interest for our research. All these demonstrates the need for the development and introduction of special courses on teaching students how to work with the tools of computer NLP instruments.

References

1. Масс-медиа КНР в условиях глобализации // СИСП. 2012. №9. С.79.
2. Международное сотрудничество в терминологических исследованиях: Сб. Статей / Под науч. ред. К.К. Васильевой, Чжен Шупу. -Чита: Поиск, 2010.
3. Мишанкина Н. А. Базы данных в лингвистических исследованиях // Вопросы лексикографии. 2013. №1. С.25-33.
4. Нагель О. В. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении // Язык и культура. 2008. №4. С.53-59.
5. Очиров О.Р. Лингвистические проблемы экономической терминологии современного китайского языка//Ученые записки Забайкальского государственного гуманитарно-педагогического университета им. Н.Г. Чернышевского. -2009. -№ 3. -С. 138-142.
6. Очиров О.Р. Становление китайского терминоведения: традиции и современность // Вестник Российского университета дружбы народов. Серия: Лингвистика. 2013. № 4. С. 116-125.
7. Очиров О.Р. Терминология современного китайского языка // Ученые записки Забайкальского государственного гуманитарно-педагогического университета им. Н.Г. Чернышевского. -2009. -№ 3. -С. 236-238.
8. Худякова О. С. Уровни ориентирующего воздействия специфических языковых структур и единиц в китайскоязычной блогосфере // Научный диалог. 2012. №3. С.138-160.
9. Чешуин С. А. Совершенствование профессиональной подготовки специалистов по лингвистике и межкультурной коммуникации на основе применения локальных вычислительных сетей // Армия и общество. 2009. №2. С.82-88.
10. Яцко В.А. Алгоритмы и программы автоматической обработки текста // Вестник ИГЛУ. 2012. №17.
11. Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank?. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics — Volume 1 (ACL '03), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 439-446.
12. Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. "A Conditional Random Field Word Segmenter." In Fourth SIGHAN Workshop on Chinese Language Processing. 2005.
13. Pi-Chuan Chang, Michel Gally and Christopher Manning. "Optimizing Chinese Word Segmentation for Machine Translation Performance" In ACL 2008 Third Workshop on Statistical Machine Translation.

Извлечение терминологии из энциклопедии Baidu

Булат Фаткулин[1,2]

[1] Южно-Уральский Государственный Университет, Челябинск, Россия
[2] Челябинский Государственный Университет, Челябинск, Россия
bfatkulin@gmail.com

Аннотация В статье рассматривается применение лингвистических технологий для преподавания ориенталистики в системе высшего образования Российской Федерации. Исследование посвящено методам извлечения терминологических единиц с использованием текстов об Афганистане на китайском языке. Приводятся решения для интенсивного автореферирования китайских текстов и предложены методы обучения студентов работе со вспомогательным программным обеспечением. Для сегментирования и извлечения именованных сущностей из текста на китайском языке использован пакет Stanford NLP.

Ключевые слова: извлечение терминологии, ориенталистика, обработка естественного языка.

Chatbot for IT Security Training: Using Motivational Interviewing to Improve Security Behaviour

Iwan Gulenko

Technical University of Munich, Munich, Germany
ivrosh@ymail.com

Abstract. We conduct a pre-study with 25 participants on Mechanical Turk to find out which security behavioural problems are most important for online users. These questions are based on motivational interviewing (MI), an evidence-based treatment methodology that enables to train people about different kinds of behavioural changes. Based on that the chatbot is developed using Artificial Intelligence Markup Language (AIML). The chatbot is trained to speak about three topics: passwords, privacy and secure browsing. These three topics were 'most-wanted' by the users of the pre-study. With the chatbot three training sessions with people are conducted.

Keywords: IT-security education, chatbots, Artificial Intelligence Markup Language, natural language processing.

1 Introduction

We strongly believe that one should refrain from stress users with education about security behaviour, if there is a technical solution. As long as there is no technical solution, security training is a necessary evil and has its place both in research and practice.

Motivational interviewing (MI) is an evidence-based treatment methodology that enables to train people about different kinds of behavioural changes [1]. It assumes that humans are willing to change for the better but often they are not capable to do so; the main reason is that they have conflicting thoughts about the change; they are not resistant but rather ambivalent. MI was already used in various fields. One example is e-therapy - smokers were able to break addiction to cigarettes when treated with MI techniques [2].

Also, chatbots were used for security education [3]: Positive attitudes of users are leveraged, when chatbots were used in an e-learning setting about security behaviour. We build on this research and combine MI with chatbots to improve users security behaviour. For this we use Artificial Intelligence Markup Language (AIML) – a basic method to simplify natural language processing[1].

[1] http://www.pandorabots.com/pandora/pics/wallaceaimltutorial.html

2 Previous work

Motivational interviewing (MI) is a way of talking to people about change. It has been used for a variety of problems including addictions, medication adherence, smoking cessation and overeating. The underlying theory assumes that often decisions are not blocked by resistance but rather by ambivalence. Often weeks or months are between knowing that a change is needed and making the change. In this period, people are in an state of ambivalence in which they want to change and do not want to change at the same time – this makes them procrastinate. MI is rooted in Self-determination theory (SDT) which is about motivation about people's growth tendencies. It presumes that people can make choices without any external influence. Research suggests that if people think they have decided to engage in a certain behaviour, they are more likely to stick to it.

The generic structure of MI can be easily adapted to security behaviour. The interviewer tries to lead the conversation from the problem, which the client wants to solve towards 'change talk' – ideas, plans and intentions to change behaviour coming from the client. In MI only this content matters, since this is most likely to be implemented by the client. A typical MI talk is a semi-structured interview divided into four phases: open questions, affirmations, reflections, and summaries. In our case asking open questions would be about typical security behaviour – for instance: *What do you do to secure computer practice or his identity online?* Instead of lecturing about some security violations, open questions enable the conversation to go into the direction of what the client really needs. The interviewer gives information or advice only when asked directly for it. Using affirmations the interviewer highlights the qualities of the client and how he managed to overcome issues in the past to engage in some desired behaviour; e.g., how the client already changed some Facebook privacy settings and is therefore capable of doing more. This should be followed by reflective listening: the interviewer tries to "guess" what the client is really thinking. It is more than just repeating what the client said; it is giving qualified guesses about what the client actually wants to say. At the end of the process the interviewer gives selective summaries of what was said. Obviously, he chooses to summarize only content that deals with 'change talk' – information coming from the client that points towards the desired behaviour [1].

Not all these facts about MI can be implemented by a chatbot. However, we believe that the first two phases, the open questions and the affirmations can be simulated by a computer. In the next section we describe in a pre-study how the first two phases of security behaviour interview MI can be conducted online.

3 Pre-study

We gathered 25 responses from Mechanical Turk (MTurk) using three basic MI questions. Through this it was feasible to get abstract information about what actually bothers internet users. The participants were paid 50 Dollar-cents per response. The survey consisted of basic questions about demographics, followed by the MI questions. We used the following wording for the questions:

– What would you like to change in your computer security behaviour?
– What hinders you to start engaging in the described behaviour?
– What would be the next steps to start engaging in the described behaviour?

We had 25 replies; the data is represented in table 1. The demographics suggest that we represent the internet user population with a small bias towards 35-54. Male and females are represented equally and the education level seems to be also representing the U.S population. In general this shows that the gathered sample has no extremes; yet we have to get more data to compare this to some bigger population. Generally, our survey results seem reliable, which confirms Buhrmeister 2011 [4]. We choose ten most substantial answers out of the 25 to give the reader a taste of the high quality of the responses (regardless of the 50 Dollar-cents payment).

The replies had mostly to do with passwords. Thirteen replies dealt with the fact that they want to improve their passwords habits. Eleven people talked about protecting their privacy online and secure browsing (protecting against malicious websites, logging out of websites, shutting down facebook. Three replies had to do with the fact that he or she wants to use better software or install an anti-virus software,. Therefore, online users (at least in our sample) mostly want to learn about passwords, privacy, secure browsing.

In the following section we develop our chatbot based on our pre-study.

4 Chatbot

The goal of a chatbot is to appear as human as possible and keep the user interested. Therefore, entertainment-wise a chatbot might be superior to traditional IT-security awareness campaigns such as posters, leaflets, mass mailings. From the viewpoint of efficancy an online trainer is much cheaper for big organizations, where the requirement is to train thousands of employees at the same time.

We develop a chatbot using pandorabots.com, a hosting platform for chatbots; it is also suits as an AIML interpreter. AIML (Artificial Intelligence Markup Language) is the state-of-the-art XML-based programming language for chatbots. Chatbots were already used in manifold contexts such as marketing, entertainment, help on smoking cessation and countless other areas. Interestingly, chatbots were also used for security education [3]: Positive attitudes of users are leveraged, when chatbots were used in an e-learning setting about security behaviour. However, Kowalski [3] does not clarify how the chatbot is programmed, which hinders us to replicate the study and forces us to build our chatbot based on questions from our pre-study, and to use MI techniques.

We briefly describe the basics of chatbots that are based on AIML. The markup language is based on XML and there are three types of tags: *patterns*, *templates* and *that*. Patterns are substrings of strings entered by the user. Patterns are nodes in a graph and edges are decisions. The chatbot traverses through a search- tree to find a path which fits the pattern. The template is at the end of a path and is the output of the chatbot. A tag called 'that' refers to the most recent output of the chatbot.

Table 1. Ten most substantial replies out of the sample of 25 participants

No.	Wanted change	Perceived hindrance	First steps
1	Protect data online, solid passwords	Lack of skills	Take class on computer security
2	Log-out of websites if not using the computer	Log-out of all websites is time-consuming	Log-out of websites after use
3	Use different passwords for different websites, figure out Facebook privacy settings.	Memorizing different passwords is hard, laziness hinders to check Facebook privacy settings	Change passwords of most-visited websites. Reading forums and Facebook FAQ.
4	Change personal data	Laziness	Take time to learn security, use antivirus for Mac
5	Use better passwords	Lack of knowledge what a good password is, many passwords	Websites should standardize password requirements
6	Buy better computer anti-virus, anti-malware, firewall to secure browsing	High cost of security software	Save money, install software
7	Use more secure browsers that do not track surfing behaviour	Inability to find such a browser	Find a browser that has a good reputation for security
8	Remember and change many passwords	Complexity of passwords (special characters, numbers)	Make complex passwords, and change suffix for different platforms
9	Change weak passwords, change them often, scan for viruses more often, change Facebook for privacy	Procrastination, nothing happened so far, virus-scan scan makes internet slow, does not care about certain logins, does not know how to adjust Facebook privacy	Acquire knowledge
10	Be less reckless, wants more protections	Money to buy security software	Get more money

We use bitlbee, an irc server, to fetch text from and send text to the chatbot. With bitlbee we can connect our chatbot to any common platform that has a chat function – e.g., Yahoo, Skype, ICQ, Facebook, Twitter. We choose Yahoo Messenger to interact with bitlbee, because of "Yahoo! Pandorabot", an open source project that seamlessly connects bitlbee with our chatbot.

We use the chat-database of *Dr. Richard S. Wallace bot 2002*. It represents a chatbot that has common knowledge that imitates a real human being. So if Dr. Wallace's bot is asked what his name is, how old he is and how he feels, he gives a reasonable answer. Additionally to this boilerplate-personality we add conversation patterns dealing with (1) passwords, (2) privacy and (3) secure browsing – exactly the requirements that we elicited in the pre-study.

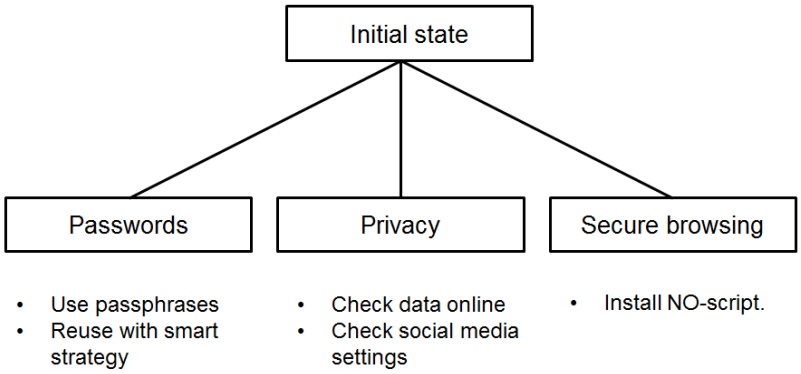


Fig. 1. Chatbot chooses the topic based on user's input and is then able to talk about different aspects of that topic

We present three sample-conversations with the chatbot. The XML pattern files that led to this conversations can be downloaded. Below we show three screenshots of the chatbot talking to clients about different topics of it-security behavioural change.

5 Conclusion

We believe that our chatbot is ideal to be used especially in big organizations, where face-to-face training of every employee is infeasible. An other way to continue the research is to use cognitive thesauri that can be used as an input for the chatbot and thus optimize its functionality. For future research, we need to test the chatbot in more usecases and how users engange into using it.

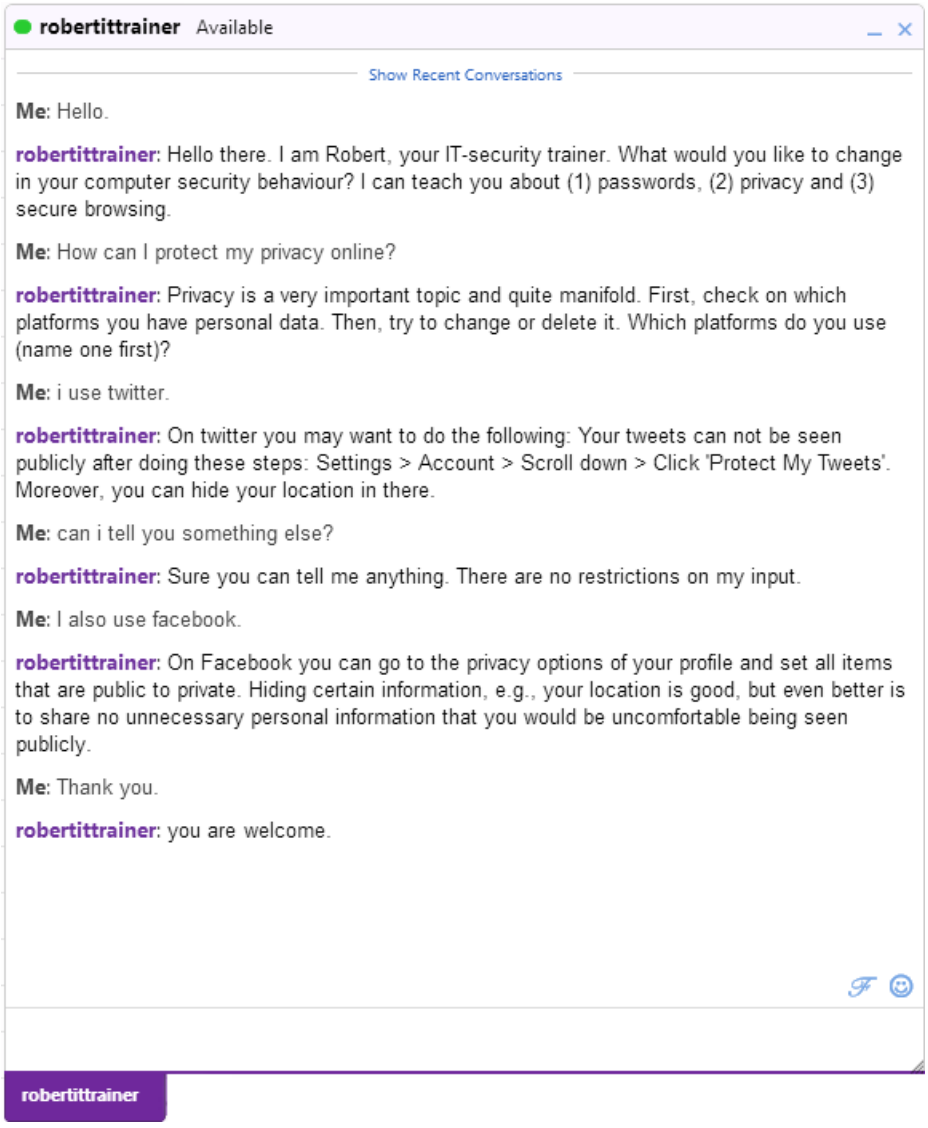


Fig. 2. Chatbot talking about privacy

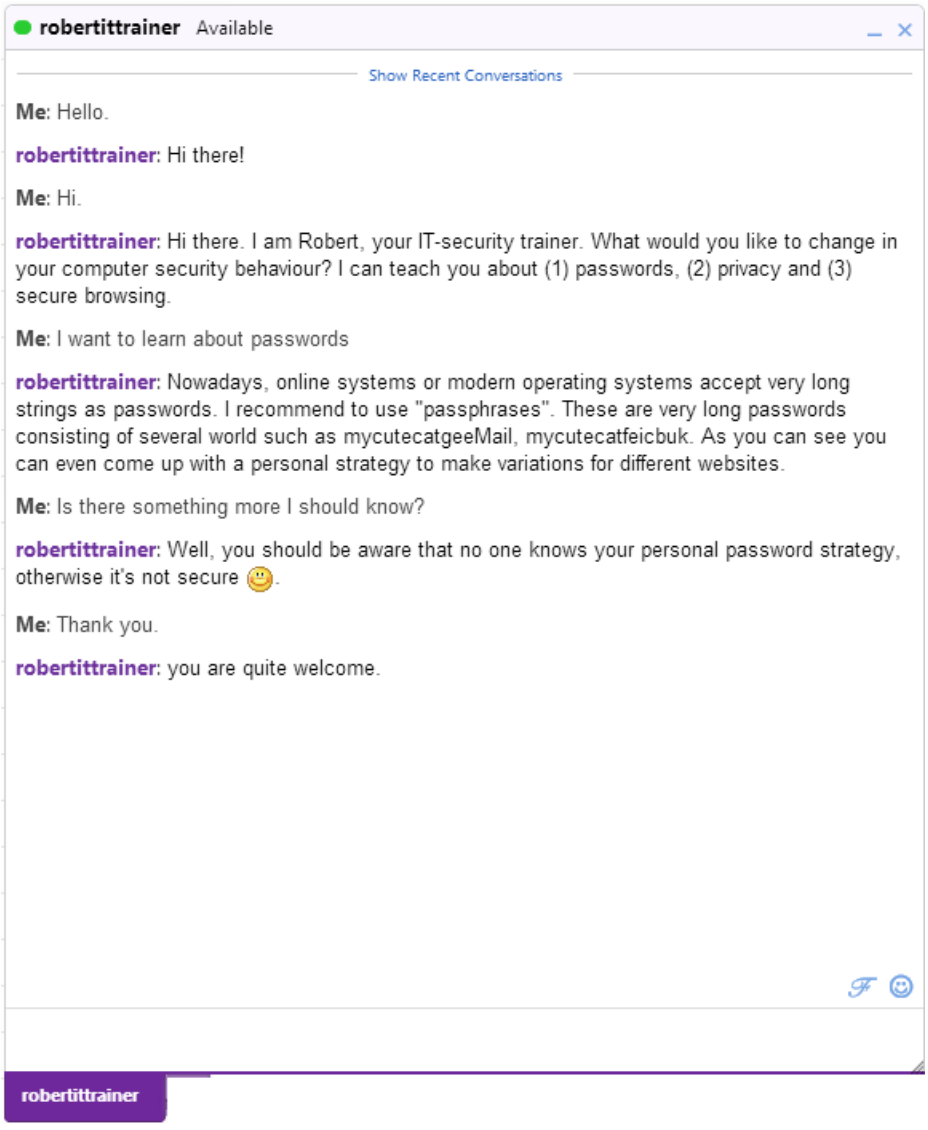


Fig. 3. Chatbot talking about passwords

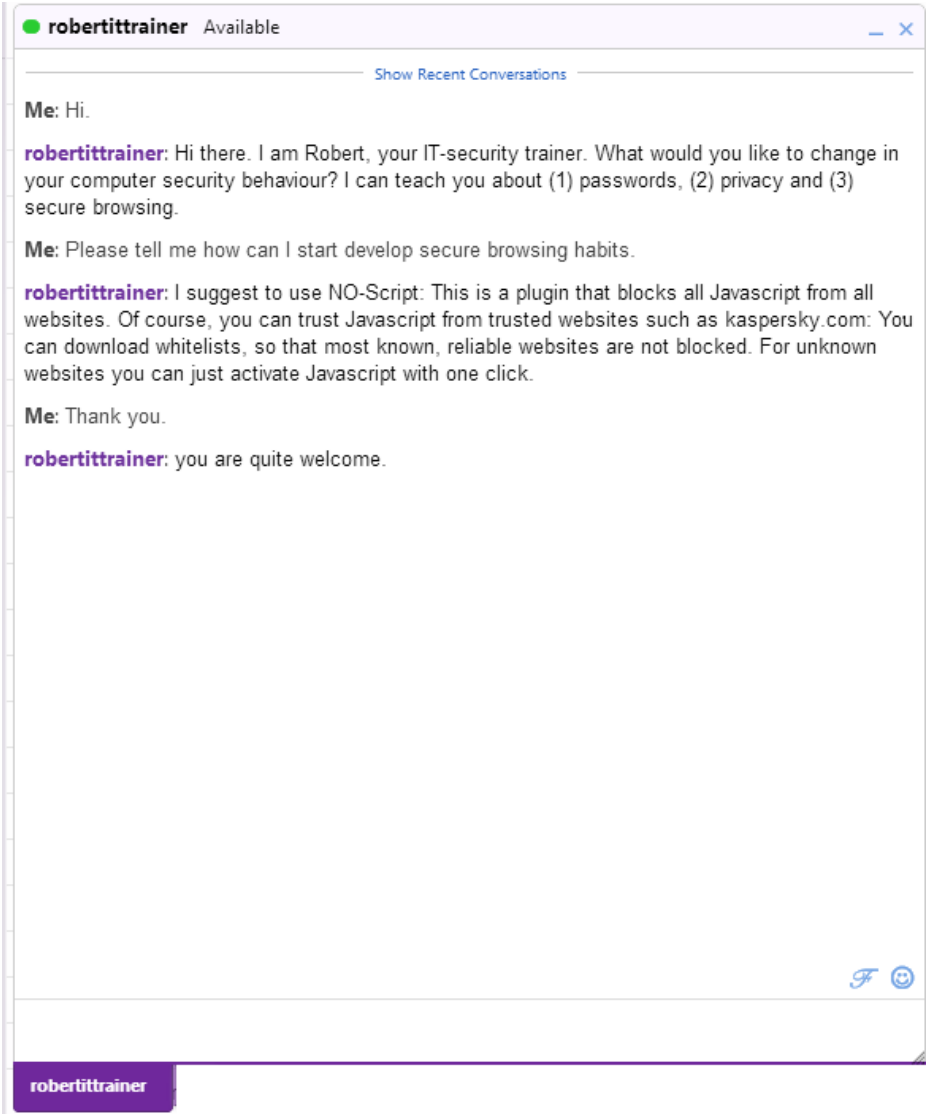


Fig. 4. Chatbot talking about secure surfing habits.

References

1. Miller, W., Rollnick, S.: Motivational Interviewing: Preparing People for Change. Applications of Motivational Interviewing Series. Guilford Press (2002)
2. Grolleman, J., van Dijk, E., Nijholt, A., van Emst, A.: Break the habit! designing an e-therapy intervention using a virtual coach in aid of smoking cessation. In IJsselsteijn, W., de Kort, Y., Midden, C., Eggen, B., van den Hoven, E., eds.: Proceedings Persuasive 2006. First International Conference on Persuasive Technology for Human Well-being. Volume 3962 of Lecture Notes in Computer Science., Berlin Heidelberg, Springer Verlag (May 2006) 133–141
3. Kowalski, S., Pavlovska, K., Goldstein, M.: Two case studies in using chatbots for security training. In Dodge, RonaldC., J., Futcher, L., eds.: Information Assurance and Security Education and Training. Volume 406 of IFIP Advances in Information and Communication Technology. Springer Berlin Heidelberg (2013) 265–272
4. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science **6**(1) (2011) 3–5

Чат-бот для обучения ИТ-безопасности: применение мотивационного интервью для повышения осведомлённости

Иван Гуленко

Мюнхенский технический университет, Мюнхен, Германия
ivrosh@ymail.com

Аннотация При помощи Mechanical Turk проведено предварительное исследование с целью определить какие поведенческие проблемы информационной безопасности наиболее важны для пользователей Интернета. Вопросы были построены в форме мотивационного интервью, позволяющего обучать людей различным формам изменяющегося поведения. На основе этого был разработан чат-бот с использованием Artificial Intelligence Markup Language (AIML). Чат-бот обучен общаться на три темы: пароли, конфиденциальность информации, безопасной просмотр Сети. По материалам предварительного исследования, в котором приняли участие 25 человек, именно эти три темы являются наиболее востребованы пользователями. При помощи чат-бота проведены три обучающих сеанса.

Ключевые слова: обучение ИТ-безопасности, чат-боты, Artificial Intelligence Markup Language, обработка естественного языка.

Conceptual Scheme for Text Classification System

Nicolay Lyfenko

Russian State University for the Humanities, Moscow
lyfenkoNick@yandex.ru

Abstract. The paper describes an application of classification algorithms to the text categorization problem. Author proposes a conceptual scheme for an automatic text categorization system. This system must operate with various text representation models and data mining methods. The novelty of this system consists in advanced implementation of JSM method for automatic hypothesis generation — an original logical-combinatorial technology of data mining, which is developed in Russia by several research groups.

Keywords: text classification system, machine learning, data mining, natural language processing

1 Introduction

Due to an increasing number of text documents in digital form and the extension of a data stream in different fields of professional activities the interest in a text categorization task has essentially increased. The main goal of classifying a new text is to assign a predefined class or classes to it [1]. It is being solved with the help of the text classification system ADC (*automatic document classifier*). Our system includes: different text representation models, a number of text mining methods and some text similarity metrics.

The main goal of the system is to compare various classical text classification methods to JSM method for automatic hypothesis generation and choose the best one for a particular task [2, 3].

This research is in progress so the main purpose of this work is to build a conceptual scheme for the ADC system, develop a project scheme for ADC system and represent its current state of work.

There is a great variety of machine learning methods to make a text classification. The most popular of are: *k-nearest neighbor, Rocchio classifier, neural network, decision trees, naive Bayes classifier,* and *support vector machine* [4–6]. There are not only algorithms but ready to use frameworks and IDE's for text classification problem (e.g. Rapidminer[1], Gate[2]). But none of them has the JSM method implemented.

This method was proposed by V.K. Finn at the beginning of the 1980s. The abbreviation JSM is given in honor to John Stuart Mill. The JSM method uses the Mill's idea

[1] http://rapidminer.com/products/rapidminer-studio/
[2] https://gate.ac.uk/

that common effects are more likely to have common causes. The JSM method for automatic hypothesis generation is known as an original set of logical combinatorial technologies for data mining using rules of plausible reasoning [7].

The JSM method includes three cognitive procedures: *induction, analogy, abduction* [2] and two main stages: *learning* (to identify data patterns using Mill's agreement) and *prediction*. By means of *induction* the JSM method generates casual hypotheses. With the help of *analogy* additional definition to unknown examples is formed (prediction). The *abduction* procedure evaluates the plausibility of the generated hypothesis.

This logical-combinatorial method for intelligent data analysis has shown good results on level with *SVM* method in the work [8] for the task of sentiment analysis. So we have a proposal to apply it in the task of automatic topic and authorship classification.

2 Conceptual Scheme for ADC System

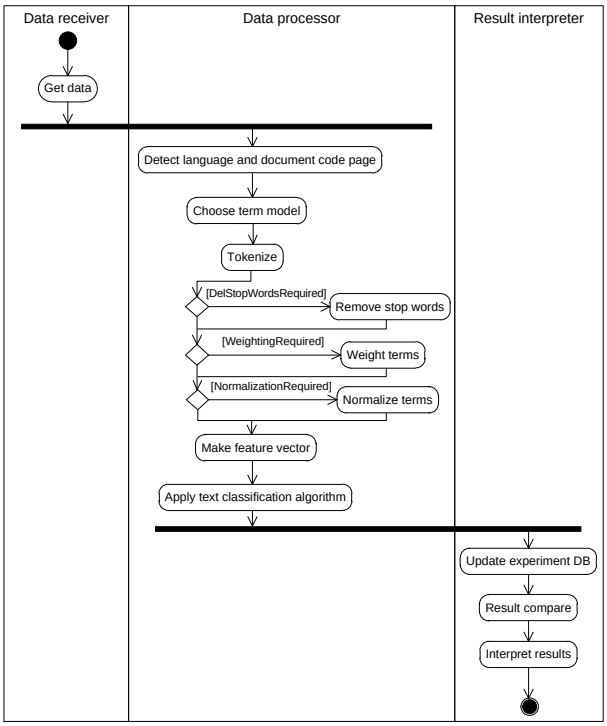


Fig. 1. Conceptual scheme for ADC system

Fig. 1 shows the key steps for automatic document classification used in the ADC system: to get data, to process it and to analyze results.

Reasoning from the fact that a document to analyze can be written in different code pages and various languages (Russian and English currently supported) a character set

and a text language should be identified. We are using statistical analysis as in [9]. In our research we normalize terms with the help of a made inverse dictionary based on Zaliznak's for the Russian language[3]. English words are stemmed.

We use some classical IR text models: *frequent* model, *tf-idf* model for text representation as an *n*-dimensional vector (*vector space model*) and not so popular but promising ones are investigated: *LOWBOW* (Locally Weighted Bag of Words Framework) [9], *MFS* (Maximal Frequent Sequences) [6], *Document Occurrence Representation* (DOR) & *Term Co-occurrence Representation* (TCOR) [9].

2.1 Project Object Model

In order to choose the best technic for a certain text classification approach we have to compare all the methods and have a log of our experiments. That is why it is proper to have well-structured and a user-friendly GUI for an experiment and logically organized project scheme for ADC system and data base for experiments.

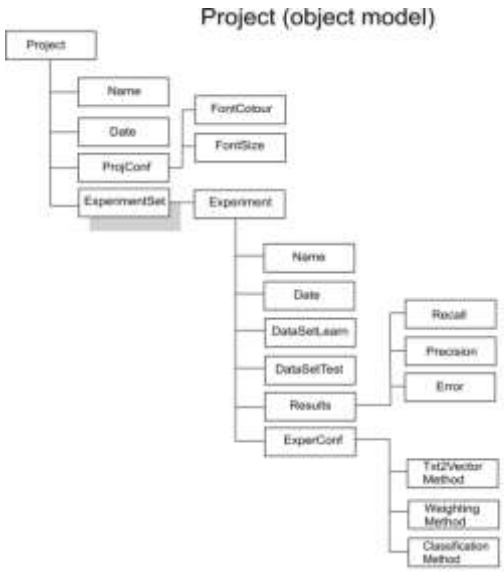


Fig. 2. Project model for ADC system

A project scheme for ADC system is represented in Fig. 2. It has *a name*, *a date* and *a project configuration* (for user's visualization preferences) properties and experiment set as a collection of experiments. It is useful to know which piece of data is used for a learning phase and a test one and what results should be shown in a log file. The property experiment configuration (*ExConfiguration*) gives the information about the text representation model, term weighting and the classification method.

[3] With the help of the COM object from www.aot.ru

3 Conclusions

In the article we suggest a conceptual scheme for an automatic document classification system (ADC). The main goal of which is to choose the best text representation model and classification algorithm for a certain application. In more detail: to compare JSM method for automatic hypothesis generation to text classification methods. That is why a project object model and its conceptual scheme are developed. The current state of the system is the following: the task of converting a text to an *n*-dimensional vector is solved. *Frequent* and *tf-idf* models for text representation are implemented. Term normalization (using the dictionary for Russian and stemming for English languages) is done.

Later the JSM method should be implemented and examined; data base scheme should be developed; experiments should be carried out and the results should be compared.

References

1. Sebastiani, F.: Machine Learning in Automated Text Categorization. J. ACM Computing Surveys vol. 34(1), pp. 1–47 (2002)
2. Finn, V.K.: Plausible inference and plausible reasoning. J. Sov Math, vol. 56(1), pp. 2201–2248 (1991)
3. Finn, V.K.: The synthesis of cognitive procedures and problem of induction. Autom Doc Math Lingust, vol. 43(3), pp.149–195 (1999)
4. Lyfenko, N.: Avtomaticheskaja Klassifikacija Tekstovyh Dokumentov na Russkom i Anglijskom Jazykah s Pomoshh'ju Metodov Mashinnogo Obuchenija. J. Molodezhnyj nauchno-tehnicheskij vestnik, vol. 4, (2013) (in Russian)
5. Cabera, J.M., Escalante, H. J., Montes-y-Gómez, M.: Distributional Term Representations for Short-Text Categorization. 14th International Conference on Text Processing and Computational Linguistics. Samos, Greece, (2013)
6. Ahonen-Myka, H.: Finding All Maximal Frequent Sequences in Text. Proceedings of the 16th International Conference of Machine Learning ICML-99 Workshop on Machine Learning in Text Data Analisys, eds. D. Mladenic and G. Grobelnik, pp.11-17, J. Stefan Institute, Ljubljana, (1999)
7. Anshakov, O.M. The JSM method: A set-theoretical explanation. Automatic Documentation and Mathematical Linguistics 46 (5),pp. 202-220,(2012)
8. Kotelnikov, E. V.: Using JSM Method for Sentiment Analysis. 3rd International Conference on Science and Technology Held by SCIEURO in London, pp.56 (2013)
9. Lebanon, G., Mao, Y., Dillon, M.: The Locally Weighted Bag of Words Framework for Document Representation. J. Machine Learning Research. vol 8, pp.2405–2441, (2007)

Концептуальная схема системы классификации текста

Николай Д. Лыфенко

Российский государственный гуманитарный университет
lyfenkoNick@yandex.ru

Аннотация. Предлагается концептуальная схема для решения задачи автоматической классификации текста. Рассматриваются различные представления текстов на естественном языке, а также статистические и логико-комбинаторные методы анализа текстов. Новизна система заключается в имплементации ДСМ метода автоматического порождения гипотез – оригинальной технологии интеллектуального анализа данных, разрабатываемой в России различными группами исследователей.

Ключевые слова. Классификация текста, машинное обучение, обработка естественного языка, интеллектуальный анализ данных.

Toward Network Information Navigation Algorithms

Sergei Bel'kov, Sergei Goldstein

Ural Federal University, Yekaterinburg, Russia,
srgb@mail.ru, vtsl@dpt.ustu.ru

Abstract. Attention is paid to the problems of automatic search of documents by search engines, analysis of documents and the use and developing of network resources, such as thesauruses and ontologies. Also some of proposals to expand the conceptual model associated with the need to reduce the dimension the set of documents found by the search engine to a set of relevant documents are formed.

Keywords: search engines, query optimization, analysis of documents.

1 Introduction

The numbers of ways we use the Internet now-a-days are really extensive. However, algorithms associated with that are almost not formalized and therefore there are many unresolved problems here. Therefore, the possibility of improving this situation, it is important.

In complex cases we have rather complicated query, and the output is the set of retrieved documents, many of which could not viewed physically, or they have duplicates of other documents or not useful for our tasks.

2 Problem of informational navigation

Problems which associated with informational navigation include there are three main components:

- Search of information, i.e. some documents or texts (books, journals, proceedings, web resources, search methods);
- Analysis of information (formats for documents, methods of obtaining the set of relevant texts, analysis methods);
- The work with network resources (dictionaries and reference books, standalone and web thesauri or ontologies).

Traditional search process (SP) of documents on the Internet can be presented by three components:

$$SP =< Q, SE, DOC >, \tag{1}$$

where Q - the set of queries; SE - many search engines; DOC - found resulting links to documents (further documents).

Query q usually includes a list of simple keywords or phrases which made up by the disjunction of conjuncts or disjunctive normal form.

Search methods which hiding within known specific search engines usually are not obvious to the user.

In addition the found resulting documents can be presented in different formats (txt, doc, pdf, ps, djvu, html, xml and others). Also a set of documents obtained by different search engines in response to the same query may vary essentially.

This raises the following tasks: selection of the most effective (in terms of search target) search engine; optimization of the structure of the query; selection from the set of received documents to only those documents that best meet the targets of the search.

Tasks associated with the optimization of the structure of the query and reducing of the set of received documents are usually beyond the capability search engines.

To resolve some of those problems we suggest to introduce feedbacks into the traditional search scheme (Fig. 1).

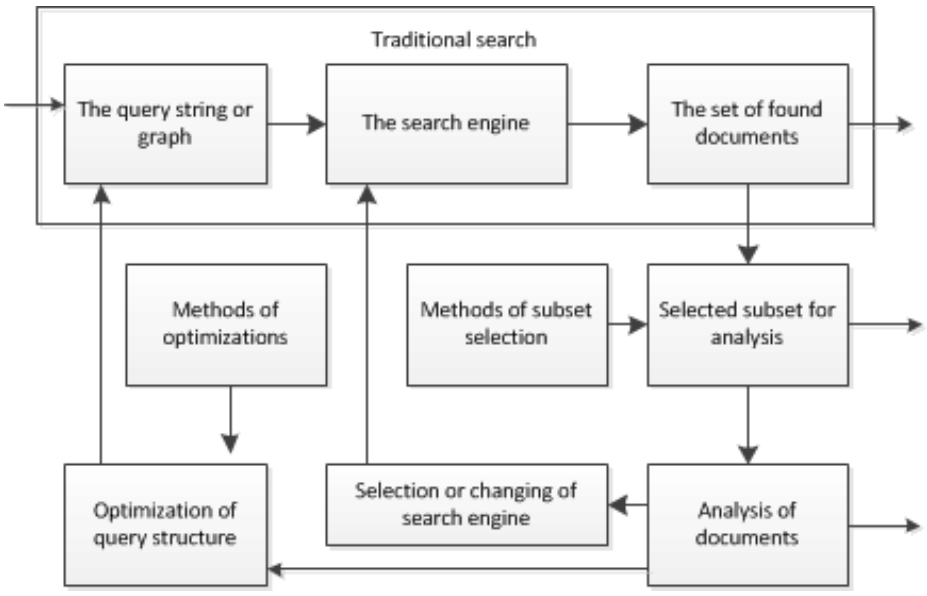


Fig. 1. Search scheme with feedbacks (the top three units are traditional)

With this in mind the procession model takes the following form:

$$SP =< Q, SE, DOC, DSM, DS, MA, A, SES, MO, QO, RDOC >, \quad (2)$$

where first three traditional components was given above; DSM is methods of subset selection for documents analysis ; DS is selection procedure of documents; MA - methods of analysis into selected subset; A is analysis procedure ; SES - selection or changing of the search engine; MO - optimization methods of query structure; QO - optimization procedure; RDOC is the resulted set of relevant documents.

Consider also some of these components separately we may suggest also some useful formalisms.

Serious problem for the analysis may be large dimension of the set of documents on the search engine output. Restricting the sample may be analyzed by random selection of documents involving experts or require the development of additional procedures.

To select a specific search engine, may write:

$$SS_k = F_{sel}(SE, C_{sel}), \tag{3}$$

where F_{sel} - select function ; SE - many of available search engines; C_{sel} - selection criteria.

The result of analysis of the set of documents obtained by applying the k-th search engine:

$$R_k = F_a(DOC_k, C_a, M_a), \tag{4}$$

where F_a - function analysis ; DOC_k - set of received documents; C_a - criteria of analysis; M_a - methods of analysis.

We also introduce the concept of optimal query:

$$Q_{opt} = F_{opt}(R_k), \tag{5}$$

where F_{opt} - optimization function of the query structure (for example graph of connections between keywords).

Often set of the found documents DOC_k is too large (typically tens of thousands). Therefore, one of the optimality criteria is to reduce the number of documents which obtained by query. Other criteria can be adequacy to the search target and complete-ness of the topic consideration.

With this in mind we may get the Algorithm of informational (text) search (Fig. 3). It is algorithm of first-level of decomposition.

At first it may demand some studies of search models or search query languages.

After that we may use for example one of the following search models: search by keys, wide primary search, random wide primary search, intellectual search, search by last heuristic, search by random walks and others types of search.

Obtained results may be divided into several groups depending on the different cri-teria or search characteristics.

Working with set of found documents will demand methods of documental analysis. During the analysis of the set of documents may appear the following tasks:

– To identify the documents which are most similar to the search aims. That may be such documents as at random taking a number of documents from the beginning of the set (for some search engines, they are usually the most relevant purpose of the request). Also more special procedures may use here (for example by taking documents one of presentation format);
– Divide the set of documents for the group (for example: unimportant, secondary importance, and high importance documents), areas or classes.

It uses a set of keywords or phrases (terms), which are presented in the documents. Some of these terms are also present in the query q. Document is describing its set of keywords is the image of the document.

For domain we have a Dictionary, consisting of terms. To determine the degree of connection between the two documents apply the mathematical apparatus of the following models: Boolean, Extended Boolean, Vectoral, Fuzzy logical, Probabilistic [1]. Nevertheless, a direct comparison of these methods is difficult, it requires the development of additional mathematical apparatus. In more complex cases, the dictionary is transformed into thesaurus or ontology. For hypertext some special form patterns may used [2].

The resulting images of documents are allowed to move to the problem of classification. There are images of reference documents (supervised learning) or clustering of documents where no master images (learning without a teacher).

The resulting matrix of pairwise proximity of documents allow us to go to their classification or clustering. Thus we gave the following tasks: exclusion of non-uninformative (in terms of search target) documents (infor-mation noise); elimination of duplicate documents; partition (classification) of the set of documents into two (important, unim-portant) or three main categories (low, medium and high degree of importance); the actual clustering as a partition of the set of documents into groups according to the properties of their images (feature vectors).

Many of the docu-ments can be excluded on the basis of viewing only its Title or Abstract. Thus there are presented three levels of consideration: primary, main and additional analysis. Turning to the image of the document as a set of keywords, we also have several levels of keywords analysis: top (from the query), middle (from Abstract) and low (from text content, i.e. known and new keywords).

Thus after analyzing the problems arising from modern network navigation we proposed to complement existing search engines several of additional units, in particular helping to optimize the structure of the query and limit the set of relevant documents.

We plan to consider them in detail in our further studies.

References

1. *Lande, D. V., Snarskii, A. A., Bezsudnov, I. V.:* Internetika: navigation in complex networks. Librokom, Moscow (2009) (in Russian).
2. *Belkov, S. A., Goldstein, S. L.:* Representation of materials of text and hypertext sources by net of patterns. J. Informational Technologies. 1 (161), p.29-34 (2010).

Алгоритмы сетевой информационной навигации

Сергей Бельков, Сергей Гольдштейн

Уральский федеральный университет, Екатеринбург, Россия,
srgb@mail.ru, vtsl@dpt.ustu.ru

Аннотация В работе представлен перечень основных компонентов информационной навигации в сети Internet. Рассмотрены вопросы оптимизации поисковых запросов и самого процесса поиска. Представлена расширенная кортежная модель поиска. Предложено несколько полезных формализмов.

Ключевые слова: поисковые машины, оптимизация запроса, анализ документов.

Semiotic System of Musical Texts

Andrew Philippovich, Irina Golubeva, Marina Danshina

Bauman Moscow State Technical University, Moscow, Russia
{aphilippovich,igolubeva,mdanshina}@it-claim.ru

Abstract. In article authors put forward a hypothesis about existence special semiotics system in music, which is close on the structure and mechanisms to a natural language. To check the hypothesis we have selected the ancient Russian chants of XI-XVII centuries, written by Znamenny notation. Using "lingvo-musical" analogies and allocation of the corresponding semiotics designs allowed applying linguistic methods to processing and analyzing chants, identification of their musical "lexicon", syntax and semantics.

Keywords: musical semiotics, Znamenny notation, computational linguistics, thesaurus, syntactic analysis, distributed-statistical analysis.

1 Musical infocognitive technologies and Znamenny chants

The research of the mechanisms of non-verbal human consciousness is one of the promising areas of infocognitive technologies. Music and related cognitive processes which are closely connected with verbal activity hold a special place in the study of these questions.

Music as well as language are a matter of communication and do not exist outside of human communication [1]. Therefore, it is always the result of some human intermediation or performance, although various natural and technological phenomena could also be the sound sources.

The hypothesis that music and language had a common ancestor – "linguomusical system" – was offered to explain the proximity of two cognitive systems. The hypothesis determined their common features [2]. During the development the systems acquired independent and unique features, but they still interact with each other.

Russian musical compositions of XI-XVII AD were recorded using special musical system (notation) which is usually called Znamenny or semiographic. It contains hundreds of special semiographic signs ("znamyas", hooks), each of them corresponds to a certain sequence of sounds of different duration and altitude. Figure 1 presents a fragment of musical manuscript in Znamenny notation.

During the time of Peter's reforms Znamenny notation was replaced by "Italian" one which was simpler and more modern linear musical system which we still use. Unfortunately, the key to decode the melodies was lost during the transformations and this doesn't allow us to translate unambiguously many ancient chants to the contemporary presentation [4].

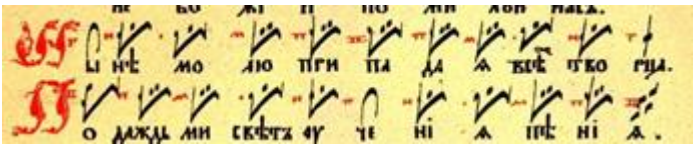


Fig. 1. A fragment of music manuscript in Znamenny notation

For the complete decryption we need to find internal laws in Znamenny notation due to which compositions contain specific signs.

While solving this task in the context of "Automated system of scientific research in the area of computer semiography" project we hypothesized that there should be some semiotic structure which is closely related by its structure and mechanisms to a natural language. This assumption allows us to apply linguistic methods to process and analyze the chants and to reveal its musical "lexicon", syntax, semantic and pragmatic.

In a case of such a hypothesis full confirmation, not merely would we possess valuable results for preserving the rich heritage of national singing culture, but also new fundamental principles of musical infocognitive technologies may be discovered.

2 Toolset development and conducting the research

To solve the problem of automated manuscript processing for selected sources, a work of several years has been carried out that included the following main stages:

- Translation of chants into a digital form;
- Carrying out basic statistical explorations;
- Informational and mathematical models development;
- Models verification, correction and application

During the first stage special computer fonts (such as "Andrew Semio") have been developed and optimized. As well, we've entered manually some semiographic chants and made necessary corrections [4].

During the second stage we've conducted statistical exploration based on an idea firstly proposed in the ancient Russian music study domain by M.V. Brazhnikov [3]. His method implies quantitative counting of semiographic signs occurrences and drawing visual graphs for subsequent analysis. We can mention a paper by B.G. Smolyakov [11] as an example of such a technology application where part one of "Dvoeznamennik 'Irmologion'"(XVII century) was analyzed with manual methods, and comparative graphs for different voices have been drawn.

First, an Andrew Tools linguistic editor was used for chants automated processing; later, a special software complex named "SemioStatistik" was developed [5-6]. It reads data in various formats (Word, Excel), parses tables and cells into constituting parts that are in their turn being written into relevant XML based data structures.

Having done some preliminary processing, SemioStatistik allows one to create various lexicographic structures such as frequency and direct concordances, vocabularies, alphabets et al. as well as to export them into different formats.

Research of dependence of frequency of a semiographic signs from its rank was conducted. It was revealed that this distribution is described by function:

$$Frec = a * e^{-b*rank},$$

a=[500;600] – depending on the manuscript
b=0,07.

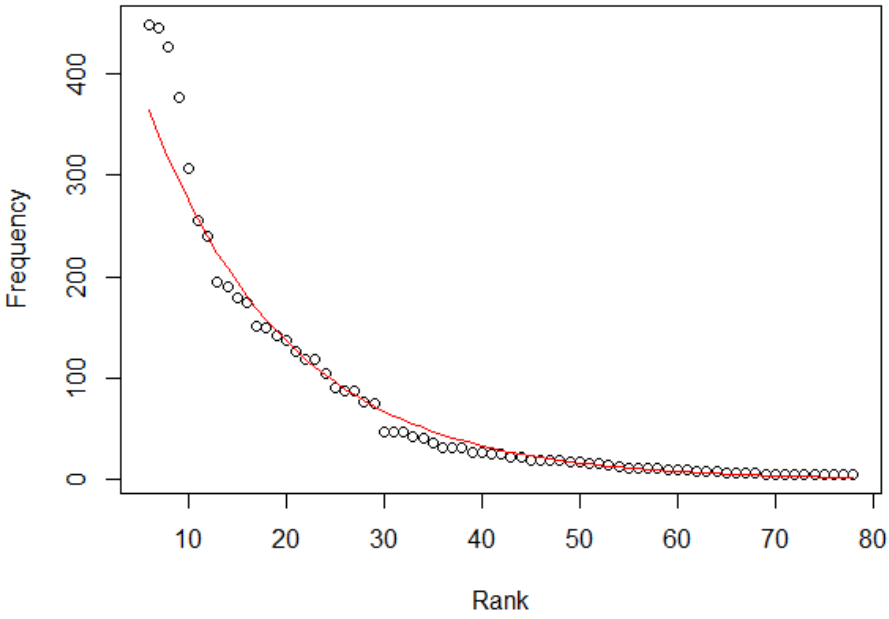


Fig. 2. Frequency dependence of the semiographic signs rank

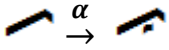
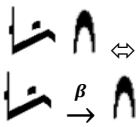
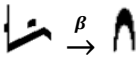
At the third stage we offered informational and mathematical models to describe components of Znamenny chants [9]. In terms of syntax we identified three types of relationships between semiographic signs (see Table 1).

The rules of semiographic signs usage could be represented in the form of Znamenny thesaurus, the structure of which includes syntax, semantic and other relationships. We present the results of building the thesaurus as dictionary entries. Every entry includes the following information: Znamya (semiographic sign), a header of dictionary entry; Basic znamya (α – relationship); Absolute frequency of znamya, numerical characteristic of znamya frequency in manuscript; Znamyas directly related to the key znamya, znamyas which go next to the key znamya (β –relationship); Znamyas context related to the key concept, znamyas which appear with the key znamya in the same context(γ – relationship).

At the fourth stage we have developed algorithms and specialized software and also have conducted a research that revealed 14 main (basic) znamyas. We could obtain other znamyas by applying the first rule (α – relationship).

To reveal syntactic relations of the second type (β-relationship) we built adjacency matrices, which contain the frequencies of znamya sequences. Further normalization concerning the overall number of znamyas allowed us to construct a stochastic table (matrix) of the transitions in Markov chain.

Table 1. Three types of relationships in the syntax of Znamenny chants

Relationship type	Description	Example
α – relationship: $Z_1 \xrightarrow{\alpha} Z_2$	Znamya Z1 is in α – relationship with znamya Z2 if Z2 is derivatived from Z1	
β –relationship: $Z_1 \xrightarrow{\beta} Z_2$	Znamya Z1 is in β – relationship with znamya Z2 if Z2 is is next to Z1	
Probabilistic β – relationship: $Z_1 \xrightarrow{\beta} Z_2 (P_i)$	If znamya Z1 could be followed by variety of znamyas then Z2 is next to Z1 the probability Pi.	 (0,56)
γ – relationship: $Z_1 \xrightarrow{\gamma} Z_2$ $Z_2 \xrightarrow{\gamma} Z_1$	Znamya Z1 is in γ – relationship with znamya Z2 if these znamyas appear in the same context (phrase, sentence, chant).	

To reveal the syntactic connections of the third type (γ-relationship) we applied a statistical distribution analysis which determined the coefficient of the «connection strength» of znamyas according to the formulas for Tanimoto metric:

$$K_{AB} = \frac{f_{AB}}{f_A + f_B - f_{AB}},$$

Figure 3 contains an example of the resulting adjacency matrix.

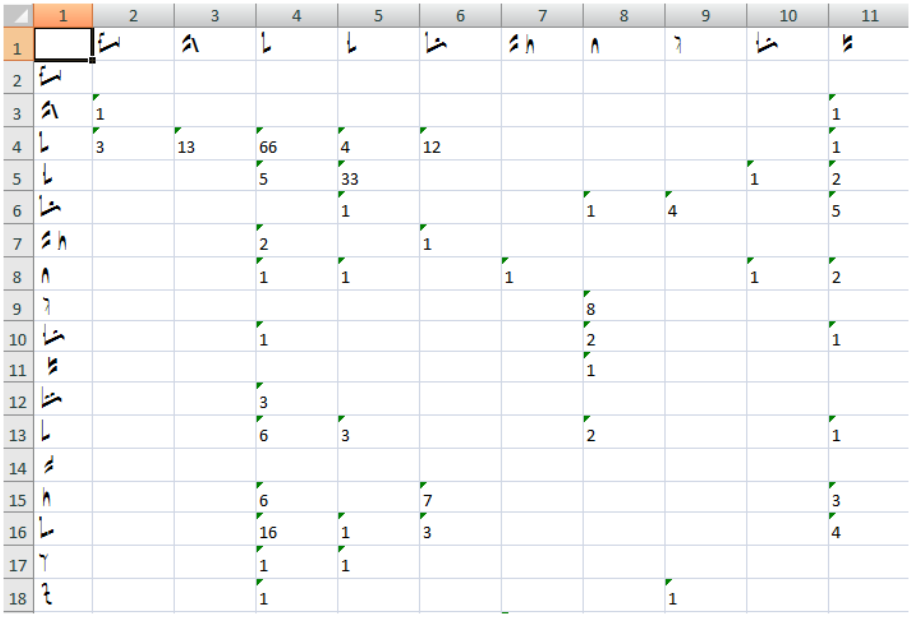


Fig. 3. Adjacency matrix (znamya sequences)

More detailed description of the model is represented in [10].

3 Conclusion

During the research we analyzed the chants from «The Ring of Ancient Znamenny Chants» book containing 24911 uses of different 722 semiographic signs (znamyas). The results of the studies support the hypothesis about the existence of complex semiotic system in Znamenny chants.

- In the general case, znamya corresponds to multiple (sequence) contemporary notes; in some cases one znamya could be replace with the group of other znamyas ("tainozamknennost"). Znamyas could be divided by typeface into main (basic) and secondary (derivative) formed by adding characteristics;
- The occurrence frequency of znamyas corresponds to the exponential law. This indicates that there is a strong spike in the probability of their usage.
- We revealed that there is a huge amount of znamya combinations that are never used; but at the same time there is a small number of combinations that are more common than others. This allows us to identify (confirm) the presence of «function» znamyas.

The application of methods of computational linguistics for the analysis of Znamenny chants, designed mathematical models and algorithms, and the results of the experiments are original and present scientific novelty in the sphere of infocognitive technologies.

The practical value of the conducted research consists of the development of software units for input, presentation and analysis of the chants, and also obtaining new statistical data about collocations of znamyas that could be used to improve data processing and to study Znamenny notation.

The obtained results provide a basis for the further studies of Znamenny chants and other musical compositions, revealing semantic and pragmatic relations, construction of new classes of personal automated systems based on infocognitive technologies.

Additional information about the project and conducted research could be found in the Internet on the website (http://it-claim.ru/semio).

References

1. Tagg P. A Short Prehistory of Western Music. Provisional course material, W310 degree course – The Institute of Popular Music, Liverpool.
2. Wallin N.L., Merker, B., Brown S. (Eds.). The Origins of Music.Cambridge, MA: MIT Press, 2000.
3. Brajnikov M.V. Ancient Russian theory of music. - Leningrad: Muzyka, 1972. (in Russian).
4. Philippovich A.Yu., Smolyakov B.G. Computational semiography // Kniga i mirovaya tsivilizatsiya: Materialy XI Mezhdunar. nauch. konf. po problemam knigovedeniya (Moskva, 20-21 apr. 2004 g.): V 4 t./ [Sost. V.I.Vasil'ev, M.A.Ermolaeva, A.Yu.Samarin; Otv. red. V.I.Vasil'ev, B.V.Lenskiy]. – M.:Nauka, 2004. — T1. – 2004. – Pp.398-401. (in Russian)
5. Philippovich A.Yu., Danshina M.V., Danshina I.V. Methods of computational semiotics on the study of ancient Russian musical writing. //XII International conference on bibliography "Biblioscience. Traditions and innovations". – Moscow: Nauka, 2009 – Pp. 359-360. (in Russian).
6. Danshina I.V. Danshina M.V. Statistical research of znamenny system based on the example of two echoes from Octoechos// Intellektual'nye tekhnologii i sistemy. Sbornik uchebno-metodicheskikh rabot i statey aspirantov i studentov. Vypusk 9. Moscow: SLC "CLAIM", 2007 – Pp. 71-80. (in Russian).
7. Smolyakov B.G. On the problem of znamenny notation deciphering. Voprosy teorii muzyki. Vypusk 3, Moscow: 1975 – Pp. 41—69.. (in Russian).
8. Danshina M.V. IPSM: A software toolset for input and processing of semiographic chants. Information technologies and written heritage: proceedings of International Scientific Conference (Ufa, October 28-31, 2010.)/ Ed. by Baranov V.A. - Ufa; Izhevsk: Vagant, 2010 – Pp. 69-74 (in Russian).
9. Philippovich A.Yu., Golubeva I.V. Syntactic research on semiographical chants // Polygraphy and publishing problems, 2012 – Pp. 147-163.. (in Russian).
10. Golubeva I.V., Philippovich A.Yu. Syntactic analysis musical texts // Analysis of Images, Social Networks, and Texts, 2013 – Pp. 196-204. (in Russian).

Семиотическая система музыкальных текстов

Андрей Филиппович, Ирина Голубева, Марина Даньшина

МГТУ им. Н. Э. Баумана, Москва, Россия
{aphilippovich,igolubeva,mdanshina}@it-claim.ru

Аннотация. Авторы статьи развивают гипотезу о существовании специальной семиотической системы в музыке, близкой по своей структуре и механизмам как естественному языку. Для проверки гипотезы взяты древнерусские песнопения XI-XVII веков, написанные в знаменной нотации. Применение «лингвомузыкальных» аналогий и размещения соответствующих семиотик позволило применить лингвистические методы для обработки и анализа песнопений, идентифицировать их музыкальный «лексикон», синтаксис и семантику.

Ключевые слова: музыкальная семиотика, знаменная нотация, компьютерная лингвистика, тезаурус, синтаксический анализ, распределённый статистический анализ.

Automatic Extraction of Hypernyms and Hyponyms from Russian Texts

Kristina Sabirova, Artem Lukanin

South Ural State University, Chelyabinsk, Russia
{bezadresa.net,artyom.lukanin}@gmail.com

Abstract. The paper describes a rule-based approach for hypernym and hyponym extraction from Russian texts. For this task we employ finite state transducers (FSTs). We developed 6 finite state transducers that encode 6 lexico-syntactic patterns, which show a good precision on Russian DBpedia: 79.5% of the matched contexts are correct.

Keywords: text mining, wordnet, hypernym, hyponym, noun.

1 Introduction

These days there is no established Russian WordNet, that is why automatic extraction of hyponyms is of great value for Russian Natural Language Processing. The concept of this work was created after the investigation of Russian syntactical structures, which contain hypernyms and hyponyms, and the review of Serelex project [14], developed for English and French [10].

The aim of this project is to extend the approach devised in the Serelex project to the Russian language. In particular, to accomplish this task we are using corpus processing tool Unitex 3.1beta [16] for automatic extraction of hypernyms and hyponyms from Russian DBpedia [12], [6]. The extracted hypernyms and hyponyms can be used to ease the creation of Russian thesauri such as RussNet [2] or YARN [3] or for query expansion in information retrieval systems.

2 Related Work

There are a lot of methods of hypernym extraction, from simple lexical patterns [7], [9], a combination of a morphological analyzer and surface syntax parsing [1], to machine learning techniques [4-5], [13] and [11]. One of the highest-coverage methods is proposed by Snow et al. [15] Firstly, they are looking for the sentences that contain two terms which are known to be in the taxonomic relations, then they parse the sentences and automatically extract patterns from the parse trees. Finally they train the hypernym classifier based on these features. Lexico-syntactic patterns are generated for each sentence relating a term to its hypernym, and a dependency parser is used to represent them.

Hearst [7] designed 6 lexico-syntactic patterns for English, which were later extended by Panchenko et al. [10] with 12 further patterns for English and French. The results of the extraction are used in Serelex, a lexico-semantic search engine. Given a query, it returns a list of related words. The system gives the opportunity to discover the meaning of words in an interactive manner, search for synonyms and more. For example, for the query "fruit" the output is "vegetable", "mango", "apple", etc. [14].

3 Russian Lexico-Syntactic Patterns for Hypernym and Hyponym Extraction

We hypothesize that the hyponymic relations are specific for most notional lexico-grammatical classes, but they are better defined for nouns and verbs. In this study, we investigate only nouns.

Our method is based on our patterns deduced in the previous work, as well as the patterns made in the Serelex project. The aim was to translate the existed patterns, to interpret them for Russian, to complete them and to create new patterns.

The Extended Abstracts corpus without accents [6] of Russian DBpedia [12] was used as the material for the research. The corpus consists of 1,325,859 sentences and ~47,000,000 tokens. For the practical part of the research Unitex was used.

Unitex is a collection of programs developed for natural language analysis using linguistic resources and tools (electronic dictionaries, grammars and lexico-grammatical tables), that gives the opportunity to develop FSTs in the graphical interface for the designed patterns. It was created for French by Maurice Gross and his students at the Laboratoire d'Automatique Documentaire et Linguistique (LADL). Similar resources were developed for other languages in the context of the RELEX laboratory network.

The electronic dictionaries specify simple and compound words with their lemmas and a set of grammatical codes. The availability of these dictionaries is the main advantage for pattern searching. The information they contain can be used for searching and matching the contexts from which the lexico-semantic relations can be extracted. These dictionaries were made by teams of linguists for different languages: English, French, Greek, Italian, German, Korean, Polish, etc. [17]. We use the full version of the Russian computational morphological dictionary, developed at CIS, Munich [8].

During the research we designed 6 patterns for the hyponym and hypernym extraction from Russian texts. For every pattern we developed a finite state transducer in Unitex and applied them to the text corpus of Russian DBPedia without accents.

To reduce the probability of matching incorrect contexts special rules were desined. These rules are mostly exceptions, enclosed in the right negative contexts to the left of the probable hypernym or hyponym. Hyponyms and hypernyms are matched using the special symbols <N> (any noun) or <!DIC> (any token not found in the dictionary). To increase the probability, that <!DIC> will match a noun, additional special symbols are placed before and after this token in the pattern, e.g. a lexico-grammatical classes like <A> for adjectives, <PREP> for prepositions, etc. or lexical masks like <первый> (first) for matching any word form of this numeral.

The patterns with the examples are presented below (X – hypernym, Y – hyponym).

Pattern 1. Такие/таких/таким X, как Y[, Y] и/или Y. (Such X as Y,[Y,] and/or Y). An example of a matched context:

В Индии зародились такие {[*религии*]=HYPER} *как* {[*индуизм*]=HYPO}, {[*буддизм*]=HYPO}, {[*сикхизм*]=HYPO} *и* {[*джайнизм*]=HYPO}.

(In India such {[*religions*]=HYPER} *as* {[*Hinduism*]=HYPO}, {[*Buddhism*]=HYPO}, {[*Sikhism*]=HYPO}, *and* {[*Jainism*]=HYPO} *were born.)*

Pattern 2. X, такие/таких/таким как Y[, Y][и/или Y]. (X, such as Y,[Y,][and/or Y]). An example of a matched context:

…{*систем* [*верований*]=HYPER}*, таких как* {[*шаманизм*]=HYPO}, {[*политеизм*]=HYPO}, {[*пантеизм*]=HYPO}, {[*анимизм*]=HYPO}.

(…{*systems of* [*faith*]=HYPER}, *such as* {[*Shamanism*]=HYPO}, {[*Polytheism*]=HYPO}, {[*Pantheism*]=HYPO}, {[*Animism*]=HYPO}*.)*

Pattern 3. X: Y[, Y] и/или Y. (X: Y,[Y,] and/or Y). A matched context:

…мир, передаваемый человеку через {*его* [*ощущения*]=HYPER}*:* {[*зрение*]=HYPO}, {[*слух*]=HYPO}, {[*обоняние*]=HYPO}, {[*осязание*]=HYPO} *и другие.*

(…the world, transferred to a human through {*his* [*senses*]=HYPER}*:* {[*vision*]=HYPO}, {[*hearing*]=HYPO}, {[*smelling*]=HYPO}, {[*feeling*]=HYPO}, *etc.)*

Pattern 4. Y[, Y][(, а также)/(также как и)/и/или] другие/другим/других/о других X. (Y,[Y,] [(as well as)/and/or] other X). An example of a matched context:

Распространение ВИЧ-инфекции связано, главным образом, с незащищенными половыми контактами, использованием зараженных вирусом {[*шприцев*]=HYPO}, {[*игл*]=HYPO} *и* {*других* {*медицинских и парамедицинских* [*инструментов*]=HYPER}}*…*

(The major modes of HIV transition are sexual intercourse, unsterile reuse of single use {[*syringes*]=HYPO}, {[*needles*]=HYPO} *and* {*other* {*medical and paramedical* [*instruments*]=HYPER}}*…)*

Pattern 5. Виды/типы/формы/разновидности/сорта X, как Y[, Y] и/или Y. (Kinds/types/forms/sorts of X, such as Y,[Y,] and/or Y). A matched context:

Такие виды {[*оружия*]=HYPER} *как* {[*шпага*]=HYPO} *и* {[*рапира*]=HYPO} *тоже причисляют к мечам, что не совсем верно.*

(Such kinds of {[*weapon*]=HYPER} *as* {[*épée*]=HYPO} *and* {[*rapier*]=HYPO} *are classified as swords, that is not totally correct.)*

Pattern 6. Y — вид/тип/форма/разновидность/сорт X. (Y is a kind/type/form/sort of X). An example of a matched context:

{[*Хобби*]=HYPO} — *вид* {*человеческой* [*деятельности*]}, *некое занятие …*
({[*Hobby*]=HYPO} *is a kind of* {*human* [*activity*]}, *some engagement, interest…)*

4 Results

We ran 6 finite state transducers for the corresponding 6 patterns on a test corpus of the first 85,071 sentences of the full corpus [6]. It contains 3,058,878 tokens. We manually verified the results, and found that 79.5% of the units were extracted correctly (see Table 1).

Table 1. The number of extracted units from the test and the full version of the corpus

Pattern	Extracted contexts from the test corpus	Extracted hypernyms (errors)	Extracted hyponyms (errors)	Errors, %	Extracted contexts from the full corpus
1	36	36 (5)	110 (12)	11.6	364
2	51	51 (13)	113 (24)	22.6	653
3	137	148 (37)	560 (120)	22.2	1402
4	97	99 (15)	284 (51)	17.2	761
5	48	48 (12)	110 (19)	19.6	395
6	59	59 (18)	59 (18)	30.5	1279
Total:	428	441 (100)	1236 (244)	20.5	4854

The second column contains the number of matched contexts with extracted hypernyms (the third column) and hyponyms (the fourth column). We also applied these 6 developed FSTs on the full corpus [6]. This yielded 4,854 extracted contexts, in which approximately 3,859 hypernyms and 11,144 hyponyms were extracted correctly.

5 Conclusion

During the research we designed 6 lexico-syntactic patterns and verified them on a large corpus. We developed 6 finite state transducers corresponding to these patterns in Unitex. These FSTs matched 428 contexts on the test corpus and 4,854 contexts on the full corpus, 79.5% of the extracted units from the test corpus were correct.

Acknowledgements. This work is partially supported by the RFH grant #13-04-12020 “New open electronic thesaurus for Russian”.

References

1. Agirre, E., Olatz. A., Arregi, X., Artola, X., Diaz de Ilarraza, A., Lersundi, M., Martinez, D., Sarasola, K., Urizar, R. Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar. In Proceedings of Euralex (2000)
2. Azarowa, I. V. RussNet as a computer lexicon for Russian (2008)
3. Braslavsky, P., Mukkin, M., Lyashevskaya, O., Bonch-Osmolovskaya, A., Krzhizhanovsky, A., Egorov, P. YARN: the beginning. In Computer Linguistics and Intelligent Technologies 2013. V. 12(19). Part 3. (2013) - Браславский, П., Мухин, М., Ляшевская, О. Н., Бонч-Осмоловская, А. А., Кржижановский, А., Егоров, П. YARN: начало // Компьютерная лингвистика и интеллектуальные технологии 2013. Т. 12(19). Ч. 3. 2013.
4. Caraballo, S. Automatic constraction of a hypernym-labeled noun hierarchy from text. In Proceedings of the 37th Annual Meeting of the ACL, pp. 120-126 (1999)
5. Dolan, W., Vanderwende, L., Richardson, S. Automatically deriving structured knowledge bases from on-line dictionaries. In Proceedings of the First Conference of the Pacific ACL, pp. 5-14 (1993)
6. Extended Abstracts Corpus without accents of Russian DBpedia: `http://cental.fltr.ucl.ac.be/team/~panchenko/data/serelex/corpus-ru-dbpedia-short-dea.csv`
7. Hearst, M. A. Automatic acquisition of hyponyms from large text corpora. In ACL, pp. 539–545 (1992)
8. Nagel, S. Formenbildung im Russischen. Formale Beschreibung und Automatisierung für das CISLEX-Wörterbuchsystem (2002)
9. Oakes, M.P. Using hearst's rules for the automatic acquisition of hyponyms for mining a Pharmaceutical corpus. In Proceedings of the RANLP Workshop, pp. 63-67 (2005)
10. Panchenko, A., Morozova, O., Naets, H.: A Semantic Similarity Measure Based on Lexico-Syntactic Patterns. In Proceedings of KONVENS 2012 (Main track: poster presentations), pp. 174-178 (2012)
11. Ritter, A., Soderland, S., Etzioni, O. What is this, anyway: Automatic hypernym discovery. In Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read, pp. 88-93 (2009)
12. Russian DBpedia: `http://wiki.dbpedia.org/Downloads39`
13. Sanfilippo, A., Poznański, V. The acquisition of lexical knowledge from combined machine-readable dictionary sources. In Proceedings of the third Conference on Applied Natural Language Processing, pp. 80-87 (1992)
14. Serelex, `http://serelex.cental.be/`
15. Snow, R., Jurafsky, D., Ng, A. Learning syntactic patterns for automatic hypernym discovery. In Proceedings of Advanced in Neural Information Processing systems, pp. 1297-1304 (2004)
16. Unitex 3.1beta. Available under LGPL license: `http://www-igm.univ-mlv.fr/~unitex/` (2013)
17. Unitex 3.1beta Manual. `http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf`

Автоматическое извлечение гиперонимов и гипонимов из русскоязычных текстов

Кристина Сабирова, Артём Луканин

Южно-Уральский государственный университет, Челябинск, Россия
{bezadresa.net,artyom.lukanin}@gmail.com

Аннотация. Описанный в статье подход по извлечению гиперонимов и гипонимов из русскоязычных текстов основан на использовании правил. Правила описаны с помощью конечных преобразователей. Мы разработали 6 конечных преобразователей, кодирующих 6 лексико-синтаксических шаблонов. Данный подход показывает достаточно высокую точность на корпусе русскоязычной DBPedia: из 79.5% найденных контекстов правильно извлечены слова, находящиеся в гиперонимических отношениях.

Ключевые слова. Анализ текста, ворднет, гипероним, гипоним, существительное.

Non-Linear Filtering of Images on the Basis of Generalized Method of Least Absolute Values

Alexander N. Tyrsin[1], Vladimir A. Surin[2]

[1] Science and Engineering Center "Reliability and Resource of Large Systems and Machines", Ural Branch of the Russian Academy of Sciences, Yekaterinburg, Russia
[2] South-Ural State University, Chelyabinsk, Russia

{at2001,sva13t}@yandex.ru

Abstract. In article consider the possibility of usage of generalized method of least absolute values for non-linear filtering of images and signals. Generalized Method of Least Absolute Values is more efficient than median methods of image processing in case of impulse interference, as well as when suppressing noise interference on high-contrast images. Workload in case of data smoothing based on Generalized Method of Least Absolute Values is comparable with the volume of calculations of median filter. Examples of realization of a method are resulted.

Keywords: generalized method of least absolute values, image filtering, median filter, impulsive disturbance, smoothing.

Introduction

Noise suppression is one of topical problems of signals and images processing. All linear filtering algorithms lead to smoothing of sharp overfalls of brightness of images under processing. This feature, being most critical in case if the information is intended for human user, shall not be conceptually excluded from the procedure of linear processing. The point is that linear procedures are optimal when Gaussian distribution takes place with signals, interference and observed data. Technically, actual images do not conform with this probability distribution. Moreover, one of basic causes of this behavior is that an image has various boundaries, brightness overfalls, texture transitions, and so on. In this respect, many real images locally described as Gaussian within the limited area, unlikely appear as Gaussian objects. This is the particular cause of poor rendering of boundaries with linear filtering.

Second feature of linear filtering is its optimality, as mentioned before, in connection with Gaussian nature of interference. Normally it is related to noise interference on images, and due to this fact, when suppressed, their linear algorithms have high rates. However, we often deal with images distorted with interference of other types. One of which is impulsive disturbance. When interference affects the image we observe white or (and) black dots randomly scattered across the frame. Application of

linear filtering in this case is inefficient, since each input pulse responds as filter pulse characteristic, and altogether they promote interference distribution throughout the frame area.

Successful solution for the described issue is the method of median filtering introduced by John Wilder Tukey [1]. Sequential processing of each point of a frame occurs when median filter is applied, resulting in formation of sequence of estimators [2]. Conceptually, processing in separate points is independent, but to speedup the process it is practical to use previous calculations on each step. Median filters are efficient when impulse noise smoothing.

But the worst case for median filtering is high-contrast image. Median filter is sensitive to high brightness overfalls. Thus, median filtering leads to signal depression, which manifests as blurred contours of contrast image details. As well, during noise suppression, pulses, which are close to each other, may persist. To eliminate the mentioned limitations, a number of various modifications to median filtering were proposed [3–5]. They may include various weighed and adaptive algorithms of medial filtering. In some cases these are of certain advantage compared to median filtering, but still they are insufficiently formalized, and normally require additional a priori information.

The research part

Let us consider the possibility of usage of generalized method of least absolute values (GMLAV) for non-linear filtering of images and signals [6].

To simplify this, we describe data smoothing with regard to signals filtering. Let us assume a non-stationary series of observations $\{x_1, x_2, ...\}$. A fundamental case of non-stationary process $x_k = a + \xi_k$ is overfall, where a is wanted signal, ξ_k – random component. In terms of data smoothing this is the study of moving filter behavior on the boundary. Behavior of moving median has been studied in many works. Therefore, let us comparatively analyze the statistic performance of moving median and GMLAV-estimators. The study will be made as per typical "overfall+noise" model [2]

$$\ldots, x_0, \ldots, x_3, x_4 + h, \ldots, x_7 + h, \ldots, \tag{1}$$

where $x_k \sim (1-\gamma)N(0,\sigma^2) + \gamma N(\mu,\sigma_1^2)$, $0 \le \gamma < 1$. Assume overfall value as (1) for certainty, as in [2], $h = 5$, and moving filter aperture as $L = 2m + 1 = 5$. In this case moving median for any number of k equals to

$$y_k^{LD} = \operatorname{med}\{x_{k-m}, \ldots, x_{k+m}\} = \operatorname{med}\{x_{k-2}, \ldots, x_{k+2}\}.$$

Moving GMLAV-estimator of mean value appears as

$$y_k^{GLD} = \arg\min_a \sum_{-m}^{m} \rho(|x_{k+i} - a|) = \arg\min_a \sum_{i=-2}^{2} \rho(|x_{k+i} - a|),$$

where ρ is a monotone increasing function twice continuously differentiable on the positive half-line, with $\rho(0) = 0$ and $\rho''(x) < 0$ for any x > 0. Let's give examples of such loss functions:

$\rho(x) = |x|^{\alpha}$, $0 < \alpha < 1$; $\rho(x) = \ln(|x| + 1)$; $\rho(x) = 1 - e^{-|x|}$; $\rho(x) = |x|/(|x| + 1)$; $\rho(x) = \arctan(|x|)$.

For certainty, let us be confined to the cases of normal distribution of random errors and symmetrical runouts. It is evident that $\forall k\ E[x_k] = 0$. Therefore, "ideally", at output the sequence shall be: … , $y_3 = 0$, $y_4 = 0$, $y_5 = 5$, $y_6 = 5$, …. Estimators of mathematical expectation (y_k) and standard deviation (s_{yk}) of mean value, median and GMLAV-statistics on sequence "boundary+noise" (1) are given in Tables 1 and 2.

Table 1. Estimators of mathematical expectation and standard deviation of mean value, median and GMLAV-statistics on sequence "boundary+noise" $x_k \sim N(0,1)$

k	Estimators	Loss functions						
		x^2	$\|x\|$	$\|x\|^{0,5}$	$\ln(\|x\|+1)$	$1-e^{-\|x\|}$	$\|x\|/(\|x\|+1)$	$\arctan(\|x\|)$
2	y_k	1,007	0,340	0,251	0,238	0,044	0,082	0,083
	s_{yk}	0,444	0,614	0,689	0,708	0,771	0,766	0,773
3	y_k	2,019	0,878	0,512	0,348	0,575	0,459	0,451
	s_{yk}	0,451	0,748	0,988	1,038	1,650	1,459	1,457
4	y_k	3,018	4,157	4,527	4,683	4,457	4,544	4,535
	s_{yk}	0,454	0,759	1,058	1,150	1,714	1,593	1,614
5	y_k	4,006	4,697	4,799	4,812	5,010	4,982	4,984
	s_{yk}	0,450	0,601	0,689	0,706	0,767	0,764	0,765

Table 2. Estimators of mathematical expectation and standard deviation of mean value, median and GMLAV-statistics on sequence "boundary+noise", $x_k \sim 0{,}9N(0,1) + 0{,}1N(0,9)$

k	Estimators	Loss functions						
		x^2	$\|x\|$	$\|x\|^{0,5}$	$\ln(\|x\|+1)$	$1-e^{-\|x\|}$	$\|x\|/(\|x\|+1)$	$\arctan(\|x\|)$
2	y_k	0,998	0,371	0,273	0,255	0,083	0,114	0,113

	s_{yk}	0,592	0,694	0,769	0,782	0,934	0,911	0,918
3	y_k	2,006	1,028	0,748	0,623	0,867	0,764	0,763
	s_{yk}	0,591	1,030	1,300	1,440	1,942	1,810	1,819
4	y_k	3,008	3,991	4,316	4,475	4,272	4,348	4,345
	s_{yk}	0,605	1,063	1,388	1,509	1,942	1,844	1,857
5	y_k	4,002	4,664	4,778	4,794	4,973	4,953	4,958
	s_{yk}	0,599	0,697	0,780	0,811	0,922	0,897	0,895

Estimations were performed with Monte-Carlo method for the number of statistical tests $M = 400000$. Estimators of mathematical expectation are displayed on Figure 1: mean value (line 1), median (line 2), GMLAV-estimator with $\rho(x) = \arctan|x|$ (line 3), and of input process at overfall (line 4) for $x_k \sim N(0,1)$.

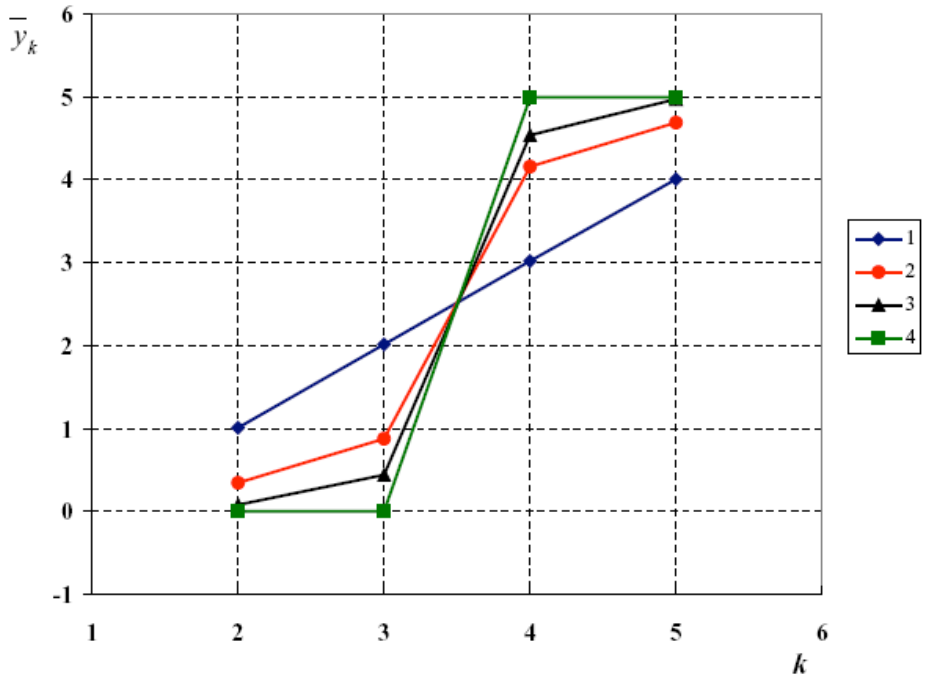


Fig. 1. Estimators of mathematical expectation of mean value (line 1), median (line 2), GMLAV-estimator (line 3), and of input process at overfall (line 4) for $x_k \sim N(0,1)$.

The analysis was performed near boundary ($2 \le k \le 5$), since with $k < 2$ and $k > 5$ mathematical expectation of all moving estimators will coincide and be unbiased. The results of the study make evident that smoothing based on GMLAV-estimators leads to less diffusion of wanted signal at overfall, as compared to median smoothing. Similar results were obtained for asymmetrical contamination.

We observe that there is a possibility of increasing the rate of Gaussian noise suppression while maintaining the same efficiency for impulsive disturbance. This is achieved by using convex-concave loss functions of type:

$$\rho(x) = \ln(|x|^{1+\delta} + 1), \qquad \rho(x) = 1 - \exp(-|x|^{1+\delta}), \qquad \rho(x) = \left[|x|/(|x|+1)\right]^{1+\delta},$$

$$\rho(x) = \operatorname{arctg}(|x|^{1+\delta}), \ \delta > 0.$$

Implementation of moving GMLAV-smoothing is not much more complex than median smoothing. Actually, in this case we do not need to resolve linear equation systems to find nodal points since they are represented by values of input process inside moving data window, namely $x_{k-m},\ldots,x_{k+m}$. The task is reduced to their simple sorting, which may be simplified by using recurrent algorithms.

Moving GMLAV-smoothing process can be applied for suppression of noise in the shape of overshoots, and for smoothing non-steady processes. However, joint analysis of smoothed and noise components has independent significance.

Let us compare two methods of smoothing. We distort a photograph of the Moon with impulse noise of 74% density. Then we process the noisy photograph (Fig. 2a) by applying two-dimensional median filter (Fig. 2b), and then by Generalized Method of Least Absolute Values (Fig. 2c), using the same aperture-cross and same number of times.

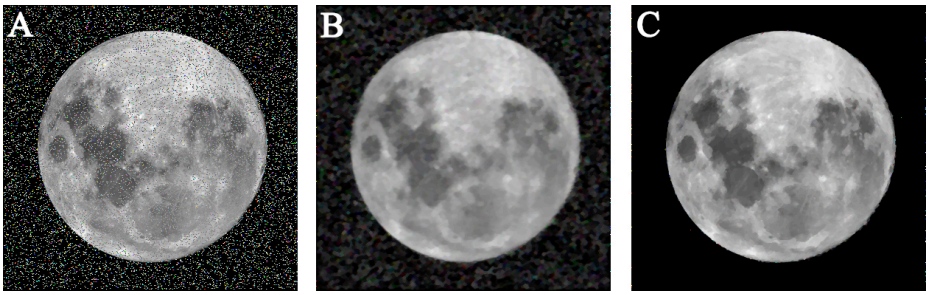


Fig. 2. Photographs: a) noisy, b) processed with median filter, c) processed with GMLAV-filter.

Signal-to-noise ratio is used as an objective criterion for image recovery quality [7]

$$W = 20\lg\frac{255}{\sqrt{s}},$$

where s is recovery mean-square error calculated by the formula

$$s = \frac{1}{N}\sum_{i=1}^{N}(\hat{f}_i - f_i)^2,$$

N is number of pixels in the image, $f(x,y)$ is initial (without additive noise) image, $\hat{f}(x,y)$ is recovered image.

For median filter we obtained WLD = 12.86 dB, at that not only noise pixels remained, but also the filter has modified the pixels of the initial image. When processed with GMLAV-filter, we obtained WGLD = 38.67 dB. These results represent higher stability of Generalized Method of Least Absolute Values against impulse interference, as well as its efficiency with regard to noise suppression of high-contrast images.

Conclusions

Thus, it is possible to draw the following conclusions:

1. Generalized Method of Least Absolute Values is more efficient than median methods of image processing in case of impulse interference, as well as when suppressing noise interference on high-contrast images.
2. Workload in case of data smoothing based on Generalized Method of Least Absolute Values is comparable with the volume of calculations of median filter.

References

1. Tukey JW (1961) Discussion emphasizing the connection between analysis of variance and spectrum analysis // Technometrics, V.3
2. Huang TS, Ed (1981) Two-Dimensional Digital Signal Processing II: Transforms and Median Filters. Berlin: Springer-Verlag
3. Abreu E, Lightstone M, Mitra S, Arakawa K (1996) A new efficient approach for the removal of impulse noise from highly corrupted images // IEEE Trans, on Image Processing, V.5
4. Chan R, Ho C, Nikolova M (2004) Convergence of Newton's Method for a Minimization Problem in Impulse Noise Removal // J. Comput. Math, V.2
5. Schulte S, Nachtegael M, De Witte V, Van der Weken D, Kerre E (2006) A fuzzy impulse noise detection and reduction method // IEEE Trans, on image processing, V.15
6. Tyrsin AN (2006) Robust construction of regression models based on the generalized least absolute deviations method // Journal of Mathematical Sciences, V.328.(In Russian)
7. Bukhtoyarov SS, Priorov AL, Apalkov IV, Khryashchev VV (2006) Application switching median filters for the restoration of noisy images. – Questions of radio electronics: Series all-technical, V.2. (In Russian)

Нелинейная фильтрация изображений и сигналов на основе обобщенного метода наименьших модулей

Александр Н. Тырсин [1], Владимир А. Сурин[2]

[1] Научно-инженерный центр «Надежность и ресурс больших систем и машин» УрО РАН, Екатеринбург, Россия
[2] Южно-Уральский государственный университет, Челябинск, Россия

{at2001,sva13t}@yandex.ru

Аннотация. Целью статьи является анализ применения сглаживания на основе обобщенного метода наименьших модулей. В ходе исследования было выявлено, что предложенный метод сглаживания приводит к меньшему растеканию полезного сигнала при наличии импульсных помех, а также при подавлении шумовых помех на контрастных изображениях. Трудоемкость реализации сглаживания данных на основе обобщенного метода наименьших модулей соизмерима с вычислительными затратами медианного фильтра.

Ключевые слова. Нелинейная фильтрация, изображение, сигнал, метод наименьших модулей, помеха, подавление помех.

Comparison of Some Image Quality Approaches

Boris B. Parfenenkov, Maksim A. Panachev

Ural Federal University
{idlerboris,tiopox}@gmail.com

Abstract. This paper is devoted to image quality problem. We analyze advantages and disadvantages of existing methods. Classification of quality metrics into some groups has been done. Based on this classification, we formed proposition about prospects of using this methods in solving image quality problem.

Keywords: image processing, image fidelity, image quality, MSE, SSIM, VDP, anisotropic.

Introduction

The problem of quality assessment arises in many different subjects. From computer graphics, where rendering of complex scenes may had a lot of time. To bioinformatics and computer security, where quality and accuracy of images may safe human lives. Although, count of image editors, which must define image quality and may improve this, significantly increases.

Certainly, our eyes is good classifier of image fidelity, but there are a lot of different software systems, which must define quality of digital images. Consequently, using of human resource is not acceptable for this problem.

In this paper we are describe of existing methods of quality image assessment. Classification of quality metrics into some groups has been done. Although analyze of advantages and disadvantages of the most promising methods was performed. We tested these metrics on the collected set of images and selected metric, which we recommend to use for solving similar tasks.

Image quality assessment

Let there be two digital images: X – original, Y – test (distorted image – with possible defects). The challenge is to build algorithm, which have these two images and define quality assessment of test image.

Digital image may be represent with brightness matrix $I = \left(p_{i,j}\right)_{H \times W}$, where $p_{i,j} \in [a, b] \cap \mathbb{Z}$, $H\ and\ W$ – height and width of image, respectively. Although, in some cases, we will consider image as one-dimensional signal $X = \left(x_{i \cdot W + j}\right)_{H \cdot W}$.

Basic metrics

At the beginning, we consider classic metrics that came to computer vision from mathematical statistics. Mean squared error of images X and Y presented as $MSE(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$. Sometimes researchers considered root mean squared error $RMSE(X,Y) = \sqrt{MSE(X,Y)}$, which may be generalize on l_p: $d_p(X,Y) = (\sum_{i=1}^{N}|x_i - y_i|^p)^{\frac{1}{p}}$.

Peak signal-to-noise ratio is calculated based on MSE and often apply for measure of distortion when image was compressed. $PSNR = 10\log_{10}\frac{L^2}{MSE}$, where $L = (b - a) -$ is the dynamic range of allowable image pixel intensities (e.g., for image that have allocations of 8 b/pixel of gray-scale, $L = 255$).

However, these metrics are not best instruments for quality assessment of images [3], because they ignore features of human image perception.

Structural similarity metrics

In paper [7] was discussed reasons of creating metrics based on structural similarity. The main idea is that human able to extract some structure from image and perceive it, but not separately pixels. Therefore, metric, which can be measure amount and kind of structural information from image, can significantly increases image quality assessment.

The first result in this approach was metric SSIM (Structural SIMilarity), which computing as composition of: illumination ($l(X,Y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$), contrast ($c(X,Y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$) and structural comparison ($s(X,Y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$); where μ_x – expected value of brightness, σ_x – standard deviation, σ_{xy} – covariance of x and y, and C_1, C_2, C_3 – some constants, that obtained experimentally.

$$SSIM(X,Y) = l(X,Y) * c(X,Y) * s(X,Y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + \mathrm{C}_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}.$$

For improve this metric we can compute weighted mean value of SSIM on local features (local feature is small part of image, which focuses people's attention; it is known, that using local features allow to discard noise and improve quality of metric [6]). Weight indicates the significance of this local feature.

MS-SSIM (Multi-scale SSIM) [8] allows to improve the image quality assessment. This metric used setp by step computation $c(X,Y)$ and $s(X,Y)$ for different resolutions. By using computation MS-SSIM for each local features we can get more impressive results [2].

SCSSIM [1] allows to evaluate the image quality by using correction for the structural features of original and test images.

Visible Differences Predictor

First this metric was developed by Scott Daly in his paper [10]. He analyzed, how to construct Human Visual System model for improving existing methods of image quality assessment.

VDP receives 2 input images and generates output differences map (each pixel has value, which describes how different the pixels of corresponding input images). Schema of work VDP presents on Fig. 1. One of the advantages of this model is pos-

sibility to get prediction of local differences between images (on the pixel level), while methods described previously provided a single value for the entire image. Although, the basis of this metric are components that are already recognized in the computer vision: CSF [11], Cortex transform [4, 11], psychometric Weibull function [9]. One of the disadvantages of VDP is non-use information about color, and work only with brightness.

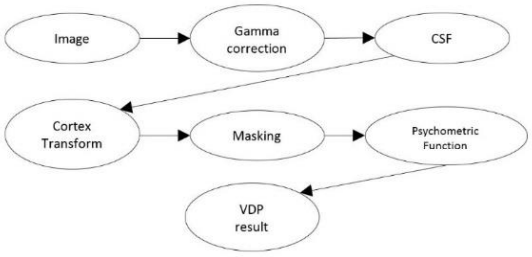


Fig. 1. Schema of Visual Differences Predictor

Anisotropy

Metrics, which described above, based on original image, and quality of test image is defined in relation to it. However, there are a variety of tasks, in which we don't have original image. Such tasks are appear in different research areas related to image quality assessment in real time. For example quality assessment of rendering complex scene in which we don't have template of image, or quality assessment of photograph that made by medical device for analyze reliability of the data, and etc.

Quality and entropy are related subjects, however, noise and information of image cannot be separated from each other (noise also has some information). For example, human with good eyesight can easily distinguish a clear object even when the image is noisy. However, analytically, entropy increases with sharpness but, in general, there is not a fair correlation when images are noisy. Hence entropy by itself is not a good indicator of image quality. And in paper [5] metric based on anisotropy was proposed, which can be represent by following:

$$LMQ(X) = B * log_{10}A(X),$$

where $A(X)$ – anisotropy of image and B is a constant that must be determined to fix the range of operative values (experimentally good results were obtained for $B = 20$).

Analyze image fidelity metrics

For analyze image fidelity metrics we collect set of images, which has been obtained by various deformations of original image. Then these images were sorted by decrease quality in terms of human perception (this sort was done using quality image assessment by three independent people). Result set of images in the sort order present on Fig. 2 (presented reduced copies, real images have a size 256x256 pixels). Names of the images correspond to deformation types: 1) shift($\boldsymbol{n}$) – shift brightness of all pixels by $\boldsymbol{n}$ (i.e. image becomes lighter); 2) noise($\boldsymbol{n}$) – add Gaussian noise (the

more $\boldsymbol{n}$ – the more noise); 3) jpeg($\boldsymbol{n}$) – image after compression by JPEG for different sizes of block (the less $\boldsymbol{n}$ – the more defects); 4) blur($\boldsymbol{n}$) – Gaussian blur (the more $\boldsymbol{n}$, the more blur).

Fig. 2. Images for testing image fidelity metrics

Then testing of 4 metrics (PSNR, LMQ, variation of SSIM, VDP) was conducted. Each metric returns evaluation image deformation – i.e. result that inverse to quality and takes values from 0 (quality image) to 1 (deformed image). The resulting graph of the image deformation (vertical axis) of the image (horizontal axis) presented on Fig. 3. Highline that main criteria for correctness of metric is not absolute deformation value but result graph has been directed upwards (i.e. for any pair of images, deformation value of left image must be greater than deformation value of right image).

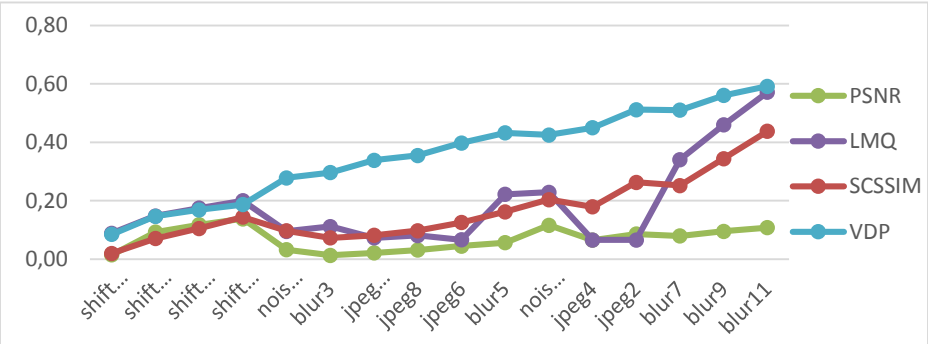


Fig. 3. Graph for image quality assessment by metrics: PSNR, LMQ, SCSSIM, VDP.

As can be seen from the graph, the most correct results were obtained by metrics VDP and SCSSIM. Metric LMQ also gave a good result, but poorly handled with images, which are compressed using JPEG.

Conclusions

In this paper we considered principal image fidelity metrics, from basic that came to computer vision from mathematical statistics, to hybrid models that use modern knowledge of computer vision and information quality assessment. Described metrics split into the following groups: statistics (MSE, l_p norm, PSNR); structural similarity (SSIM, MSSIM, MS-SSIM, IW-SSIM, SCSSIM); human perception (VDP); anisotropy (LMQ). Advantages and disadvantages of these methods were considered.

In the result of this work we determine that the presence of the original image should be use VDP metric. If we don't have original image for quality assessment, then should be use LMQ, which is very costly from a computational point of view.

References

1. K. Gu, G. Zhai, X. Yang, and W. Zhang "An improved fullreference image quality metric based on structure compensation", APSIPA ASC, pp. 1 - 6, 2012
2. Z. Wang and Qiang Li "Information content weighting for perceptual image quality assessment", IEEE Transaction on Image Processing, vol. 20, no. 5, pp. 1185-1198, 2011
3. Z. Wang and A. C. Bovik "Mean squared error: Love it or leave it?-A new look at signal fidelity measures," IEEE Signal Processing Mag., vol. 26, no. 1, pp. 98-117, 2009
4. A. S. Lukin. "Improved Visible Differences Predictor Using a Complex Cortex Transform", 19-th International Conference on Computer Graphics GraphiCon'2009, pp. 145−150, 2009
5. Gabarda, Salvador, and Gabriel Cristóbal. "Image quality assessment through a logarithmic anisotropic measure", Photonics Europe. International Society for Optics and Photonics, pp. 70000J-70000J, 2008
6. Datta, Ritendra, Jia Li, and James Z. Wang. "Content-based image retrieval: approaches and trends of the new age", Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval, pp. 253-262, 2005.
7. Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity", Image Processing, IEEE Transactions on 13.4, pp. 600-612, 2004
8. Wang, Zhou, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale structural similarity for image quality assessment", Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on. Vol. 2, pp. 1398-1402, 2003
9. Wichmann, Felix A., and N. Jeremy Hill. "The psychometric function: I. Fitting, sampling, and goodness of fit", Perception & psychophysics 63.8, pp. 1293-1313, 2001
10. Daly, Scott J. "Visible differences predictor: an algorithm for the assessment of image fidelity", SPIE/IS&T Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics, pp. 2-15, 1993
11. Mannos, James, and David J. Sakrison. "The effects of a visual fidelity criterion of the encoding of images", Information Theory, IEEE Transactions on 20.4, pp. 525-535, 1974

Сравнение методов оценки качества изображений

Борис В. Парфененков, Максим А. Паначев

Уральский Федеральный Университет
{idlerboris,tiopox}@gmail.com

Аннотация. Рассматривается задача оценки качества цифрового изображения. Проанализированы преимущества и недостатки существующих методов. Проведена классификация метрик качества в некоторые группы. На основе анализа и тестирования методов на выбранных изображениях, сформировано утверждение о перспективности использования выбранных методов в решении поставленной задачи.

Ключевые слова. анализ изображений, метрики качества изображений, image fidelity, MSE, SSIM, VDP, анизотропия.

Zipf's Law for LiveJournal

Nikita N. Trifonov

Kazan Federal University, Kazan, Russia
nikita-trif@yandex.ru

Abstract. The paper provides an overview of research of frequency of language units on the material of the LiveJournal corpus. The corpus includes texts on Russian language from 2002 to 2014 year, totaling more than 5 million words of articles written by 2 thousand authors. Research was held in the following main directions, represented in the present work: estimation of coefficients for the Zipf's law for different authors, estimation of coefficients for the Zipf's law for the total number of words in all the analyzed articles.

Keywords: Zipf's law, LiveJournal corpus, frequency of words, rank distribution.

1 Introduction

Perhaps the most famous statistical distribution in linguistics is Zipf's law: in any large enough text, the frequency ranks (starting from the highest) of wordforms or lemmas are inversely proportional to the corresponding frequencies [1]:

$$f(r) * r = c, \tag{1}$$

where $f(r)$ is the frequency of the unit (wordform or lemma) having the rank r and c is a constant. With Mandelbrots improvements to Zipf's law, the formula (1) has next form [2]:

$$f(r) = \frac{c}{r^{\gamma}}, \tag{2}$$

where γ is the exponent coefficient (near to 1). Zipf's law is most easily observed by plotting the data on a $log - log$ graph, with the axes being $log(rank\,order)$ and $log(frequency)$. After taking the logarithm of the formula (2) :

$$ln(f(r)) = C - \gamma ln(r), \tag{3}$$

The LiveJournal source chosen to collect the corpus because it makes it possible to explore articles as a whole and separately for each author.

The LiveJournal[1] (LJ) is a social network owned by SUP Media where Internet users can keep a blog, journal or diary, and is also the name of the free and open source server software which runs the LiveJournal website and online community.

[1] http://www.livejournal.com/

In order to collect corpus of LiveJournal, created a program that gets the text of articles written by one author, saves the text in a database and goes over to another author for further information gathering.

2 Experimental Results

The graph plotted (Fig. 1) using for Zipf's law the points: $x_r = log\, r$, $y_r = log\, f(r)$ where $r = 1 \dots n$, and n is the number of different units (wordforms or lemmas). The Ordinary Least Squares used to approximate such a graph by a straight line $y = ax + b$, where a and b correspond to γ and C for Zipf's law (the formula (3)) .

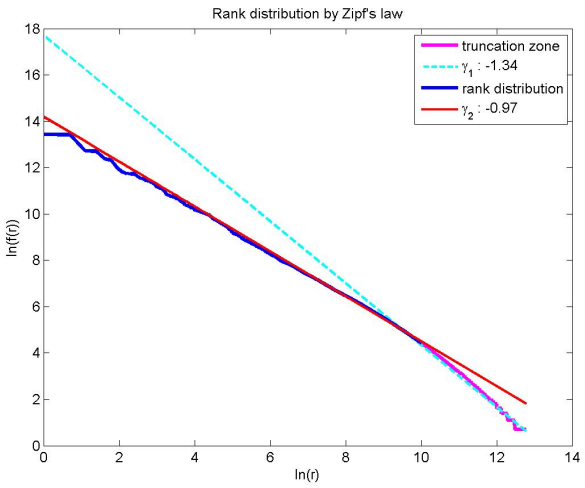


Fig. 1. The Zipf's law for the total number of words in all the analyzed articles.

The graph (Fig. 1) can be divided into three different parts. The first part is a nuclear zone consisting of the most frequently used words in the Russian language - prepositions, pronouns, etc. The central part of the graph very important for exploring. It is most accurately described by Zipf's law. The last part, called "zone of truncation " consists of words which do not carry meaning, rarely used terms and grammatical errors.

As the graph shows, the zone of truncation affects the result of the approximation, and γ coefficient in approximating line is differs from the expected. However, if we do not consider the zone of truncation, approximation line almost merges with the graph of frequency distribution, and γ coefficient satisfies the improvements of Mandelbrot for Zipf's law.

Ten authors, who written the highest number of letters in articles, were selected for further researches.

Table 1. List of 10 authors with the highest total number of words in articles

Author page on LiveJournal	The total number of words used by the author
http://eto_fake.livejournal.com/	584061
http://mzadornov.livejournal.com/	583071
http://cuamckuykot.livejournal.com/	347962
http://aillarionov.livejournal.com/	302165
http://mgsupgs.livejournal.com/	260479
http://matveychev_oleg.livejournal.com/	243579
http://steissd.livejournal.com/	240800
http://kak_eto_sdelano.livejournal.com/	234767
http://annatubten.livejournal.com/	225701
http://adamashek.livejournal.com/	214743

For each author from the list given in Table 1 made separate research. These researches have shown that all of graphs correspond to the Zipf's law. An example of this graph you can see in Figure 2.

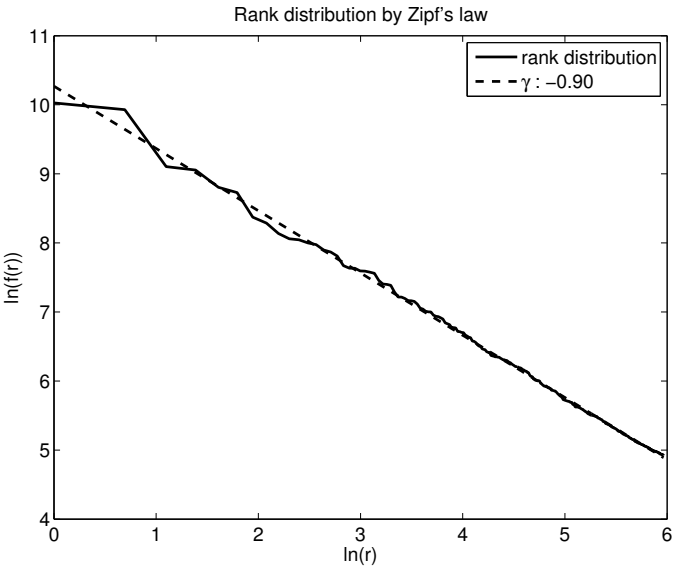


Fig. 2. Rank distribution by Zipf's law for http://eto_fake.livejournal.com/

For comparison, made the research of rank distribution of word frequencies of Zipfs law based on 4 volumes of books Leo Tolstoy's "War and Peace." (Fig. 3)

There are differences between the list of the most frequently encountered words of Leo Tolstoy's works and the list of the most frequently encountered words of contemporary authors represented on LiveJournal. These differences are

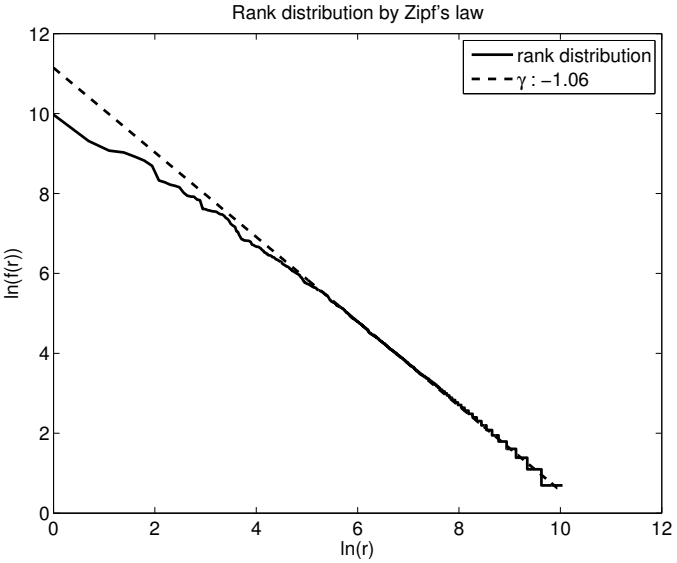


Fig. 3. Rank distribution by Zipf's law for L. N. Tolstoi

related with the difference between the vocabulary of Leo Tolstoy and modern vocabulary, as well as skills in writing texts.

3 Conclusion

Exponential coefficients of Zipf's law depend on text volume, genre of the text and author's style. The zone of truncation encountered in the research of large texts or texts written by different authors. Explanation of this phenomenon needs more investigation.

Acknowledgements. I would like to thank Valery Dmitrievich Solovyev, Eduard Yulyevich Lerner, and Vladimir Vladimirovich Bochkarev for helpful discussions.

References

1. *Zipf, G. K.* Human behavior and the principle of least effort. Cambridge, MA, Addison-Wesley, 1949, p. 36.
2. *Mandelbrot, B.* An informational theory of the statistical structure of languages, Communication Theory, ed. W. Jackson, Betterworth, 1953, pp. 486502
3. *Gelbukh, A., Sidorov, G.* Zipf and Heaps Laws' Coefficients Depend on Language. Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332–335.

Закон Ципфа для LiveJournal

Никита Н. Трифонов

Казанский (Приволжский) федеральный университет, Казань, Россия
nikita-trif@yandex.ru

Аннотация В статье представлен обзор исследований частоты языковых единиц на материале корпуса LiveJournal. Корпус включает тексты на русском языке, написанные в период с 2002 по 2014 год. Были исследованы статьи 2000 авторов, а так же более 5 000 000 словоформ из этих статей. Исследование было проведено в следующих основных направлениях, представленных в настоящей работе: расчет коэффициентов закона Ципфа по отдельным авторам, расчет коэффициентов закона Ципфа по всем проанализированным статьям без дифференциации по авторам.

Ключевые слова: закон Ципфа, корпус LiveJournal, частота встречаемости слов, ранговое распределение.

Moving Object Detection in Video Streams Received from a Moving Camera

Sergey Starkov, Maksim Lukyanchenko

National Research Nuclear University MEPhI, Obninsk, Russia
starkov@iate.obninsk.ru, maksim.lukyanchenko@gmail.com

Abstract. Detection of moving objects in a video stream received from a moving camera is difficult computer vision task, because the motion of the camera blends with the motion of the objects in the scene. In order to tackle this problem, we propose a method based on optical flow calculation and Delaunay triangulation. Given a sequence of frames, firstly, we extract the corner feature points using ORB algorithm and compute optical flow vectors at the extracted feature points. Secondly, we separate the optical flow vectors using K-Means clustering method. Third, we classify each cluster into camera and object motion using its mean scatter value. Finally, we represent the moving object using Delaunay triangulation.

Keywords: moving objects, moving camera, unstable background, ORB, optical flow, clustering, Delaunay triangulation

1 Introduction

Detection of moving objects of interest and tracking of such objects from frame to frame is an important tasks in systems that perfoms of video data, such as video surveillance systems, industrial robots, unmanned vehicles etc.

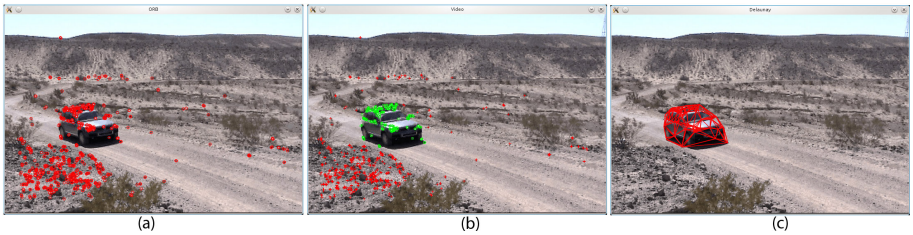


Fig. 1. Moving Object Detection in Video Streams Received from a Moving Camera (a)ORB feature points. (b)Classification of feature points, (c)Moving object in video stream

Based on the type of motion between frames, all objects can be divided into two classes: static and dynamic. Static objects maintain their position on the

sequence of frames, dynamic objects change their position in space. There has been considerable research focusing on the separation of static and dynamic objects in video sequences taken from camera in stationary positions.

The classical object detection methods can not be applied directly for detecting such objects in a scenario with a moving camera because there exist multiple sources of motions from both the camera and the moving objects. In our research we focus attention on the problem of moving object detection in a video stream captured using a moving camera. To detect moving objects in a moving camera environment we need to discriminate between camera motion and object motions. Generally, there are three approaches to detect the moving object under a moving camera:

- Compensation of camera motion by ego-motion estimation [1][2];
- Separation of motions vectors in the input sequence using motion models [3][4];
- Segmentation of the camera and object motions using the graph cut algorithm [5][6].

Some of these methods need an additional algorithmic stage to select the moving object motion model, others require considerable computation time. We decided to design an improved moving-object detection method using data from a free-moving camera with a non-stationary background, which provides both high detection performance and fast processing speed. In this paper we demonstrate the proposed approach with initial results.

2 Motion estimation

For extracting structured information about the object of interest in the image, we search for feature points of the scene.

A feature point in a scene $\boldsymbol{M}$ is called a scene point if it is coplanar with other points in image neighborhood $O(M)$ that can be distinguished from all other neighborhoods $O(N)$, which in turn are composed of several points $\boldsymbol{N}$.

2.1 Feature Point Detection

In the proposed method, to find the feature points of the image, the Oriented Fast and Brief detector (ORB) is used (Fig. 1 (a)). When using this technique, it is assumed that the intensity of the corner point is offset from the center and this displacement vector can be considered as feature point direction. To calculate a descriptor of the point $\mathbf{p}(x, y)$, ORB compares brightness values of points located in its vicinity. This algorithm is invariant to image rotation, scale change, and changes in lighting level, so it satisfies the main qualities required of robust feature detectors and is suitable for a reliable estimation of moving singular points.

2.2 Optical flow computation

To determine the movement of objects in two-dimensional space using optical sensor systems, algorithms in computer vision and image processing make use of optical flow - the apparent motion of the image, which is a shift of each point between two consecutive frames.

In our approach, we compute the optical flow vectors of image points by searching for the corresponding feature points between two consecutive image frames using the pyramidal Lucas-Kanade method. This process consists of two tasks: generation of image pyramid and search for the correspondence feature points on the image pyramid.

3 Motion Clustering

Clustering is the division of the set of input vectors into groups (clusters) on the degree of "similarity" to each other.

In this paper, we cluster feature points using the length L and direction θ of optical flow vectors. The feature points are described in the optical flow coordinate (L, θ).

All optical flow coordinate (L, θ) were divided into blocks. Randomly were selected the initial points for clustering. Number of clusters is an input parameter of the method. In the present implementation of the algorithm is assumed two: background and foreground.

4 Motion classification

The clusters generated are to be separated into those that relate to the movement of the camera and those to moving objects.

In the proposed framework, we assume that the background occupies a larger area of the frame than moving objects. Thus, the points that relate to the background have a greater dispersion than singular points belonging to objects in motion (except in cases where the background has a large amount of small details). The assignment of each cluster to a background or a moving object can be done using the measure of spread of the points within the cluster. To determine the measure of the spread of points within each cluster, in the present work, we use the standard deviation s as a discriminative metric.

Cluster, which has the highest standard deviation, be deemed to apply to the background (Fig. 1 (b)).

5 Moving object detection

To select the area that relates to a movable object, in the proposed framework, Delaunay triangulation has been used. Triangulation is a planar partition of the 2D space by plane figures, one of which is an outer infinity, and the rest are triangles.

When using Delaunay triangulation, for all resulting triangles, points of the cluster except for points at vertices lie outside the circle circumscribed about the triangle.

After constructing the Delaunay triangulation, the resulting set of triangles with edges length exceeding a predetermined threshold are removed (Fig. 1 (c)).

6 Conclusion

In this work, we have developed an effective method for separation of moving objects in the scene using data from an input video stream in the presence of a non-stationary background. This method shows high frame rate performance - 20-21 fps on a computer with a processor Intel Xeon E5420 1333 MHz and 4GB RAM. However, this value does not satisfy the operation mode in real time (>24 frames per second). In order to improve the real-time performance of the algorithm, we envisage that in subsequent implementations, in addition to algorithmic optimization, implementation will be carried out on a graphics card using software optimization libraries for CUDA and OpenCL. Also planned are:

- Exploration of the possibility of introducing additional parameters to improve the quality of clustering.
- Implementation of automatic identification of cluster numbers in the step for clustering singular points.
- Implementation of automatic identification of cluster numbers in the step for clustering singular points.

References

1. Hayman E., Eklundh J.: Statistical background subtraction for a mobile observer. In: Proc. IEEE ICCV 2003. (2003)
2. Ren Y, Chua CS, Ho YK: Statistical background modeling for non-stationary camera. Pattern Recogn **24(1-3)** (2003) 183–196
3. Borshukov GD, Bozdagi G, Altunbasak Y, Tekalp AM: Motion segmentation by multistage affine classidication. IEEE Trans Image Process **6(11)** (1997) 1591–1594
4. Ke Q, Kanade T: A subspace approach to layer extraction. In: IEEE CVRP. (2001)
5. M, X.J.S.: Motion layer extraction in the presence of occlusion using graph cuts. IEEE Trans Pattern Anal Mach Intell **27(10)** (2005) 1644–1659
6. D, S.T.C.: High resolution motion layer decomposition using dualspace graph cuts. In: In: Proc. IEEE CVPR. (2008)

Выделение движущихся объектов сцены из входящего видеопотока при наличии нестационарного фона

Сергей Старков, Максим Лукьянченко

НИЯУ МИФИ, Обнинск, Россия
starkov@iate.obninsk.ru, maksim.lukyanchenko@gmail.com

Аннотация Автоматическое выделение движущихся объектов сцены из входящего видеопотока – одна из важнейших задач анализа изображений. За последнее время было предложено большое количество методов решающих данную задачу при условии неподвижности видеокамеры. В основе этих методов лежит принцип накопления кадров и выявления изменений в них. Однако при наличии подвижной камеры применение данного подхода становится невозможным. Вместе с тем развитие автономных беспилотных транспортных средств требует решения и этой задачи компьютерного зрения. В представленной работе предложен метод нацеленный на решение указанной задачи, в его основе лежит выделение ключевых точек изображения, вычисление оптического потока, сегментация изображения на фон и объекты относящиеся к переднему плана, маркирование участков изображения. Произведена оценка результатов работы предложенного метода.

Ключевые слова: движущийся объект, подвижная камера, нестабильный фон, оптический поток, кластеризация, триангуляция.

Automatic Selection of Verbs-Markers for Segmentation Task of Process Descriptions in Natural Language Texts

Varvara A. Krayvanova

Altai State Technical University, Barnaul, Russia
krayvanova@yandex.ru

Abstract. The paper presents the intermediate results of the research, the final goal of which is to develop the universal algorithm for process diagrams automatic visualization by text description of these processes. The purpose of this study is to check the use of verbs as markers for the semantic labeling of long fragments in scientific texts.

Keywords: automatic text fragmentation, text mining of scientific texts, verbs-markers, dynamic text parameters.

1 Introduction and Problem Statement

An effective system of collecting, storage and data processing of scientific observations will raise any natural-science research to essentially new level. The description of objects and procedures is presented in the form of natural language texts. Therefore automatic transformation of texts into more effective representations (such as activity diagrams, state diagrams, use case diagrams, IDEF0 diagrams, etc.) is required to reduce the cost of specialized information systems creation.

Current researches in the field of process descriptions extraction from natural language texts are oriented to work with the news bulletins [1] or with other objects from a very narrow areas [2,3]. The similar situation is beheld with the problem of process visualization [4,5]. These algorithms assume the texts of the small length containing concentrated information of a certain type. Researches in the field of text processing for arbitrary structure and size are usually oriented to extraction of objects, instead of processes, for example, on ontologies construction[6]. To generalize existing algorithms of processes extraction to the long natural language texts we have to use automatic segmentation and semantic marking of the text to find places, suitable for these algorithms usage. The objects of this research include long scientific, regulatory and educational texts (articles, tutorials, monographs).

To reach these goals, it is necessary to allocate text fragments with various assignments:

– static (descriptions of objects, definitions) for ontology extraction;

– dynamic (description of processes, techniques and research procedures) for activity diagrams and other process diagrams extraction.

A text can be divided into various fragments of these types with the use of clustering on the base of statistical analysis of parts of speech distribution[7]. We simulated the reading process by the sliding window method (window is the sequence of the length L of consecutive sentences). As clustering parameters, for each window the total number of words and the number of various words separately for nouns, verbs and adjectives are calculated. The studies of the various parameters distribution in long texts are focused mainly on the definition of the author [8]. This method allows to divide scientific texts into fragments of the types described above. For automatic illustration we have to find a way to define fragments types. One possible way of solving this problem is to analyze the distribution of verbs in the text. There are usually much less various verbs than nouns in the texts, especially in business and scientific ones[9]. Linguistic verbs classifications, e.g. the one given in the dictionary of linguistic terms by Rosenthal[10], are not good enough for extracting information from scientific texts. In scientific style, quite narrow verbs segments are applied, therefore linguistic classifications can be called excessive. Besides, used verbs and their meaning in the text significantly depend on concrete subject domain.

The purpose of the research is to check the possibility of using verbs as markers for different types of fragments.

For the illustrations we used *Bykov N. I., Popov E. S.* Observing the dynamics of snow cover in protected areas of the Altai-Sayan Ecoregion. Methodological guidance. Krasnoyarsk. 2011. 64 pages. The text consists of 1257 sentences. Parser identified 196 different verbs.

2 Mathematical model

Let V be the set of all natural language verbs. Scientific text T is represented as an ordered set of natural language sentences $T = \langle s_k \rangle$, where s_k is kth sentence in the text. Let $V_k \subset V$ be the set of verbs in the sentence s_k. For each verb let's define the list $E_v = \langle s_k | v \in V_k \rangle$. This is an ordered list of sentences that contain a verb v. $|E_v|$ is the number of occurrences of the verb v in the text T. Since the object of study is the verb distribution in the text, the cases of multiple use of a single verb within a sentence can not be ignored. Text neighborhood $T_v^\epsilon = \langle s_i \rangle$ of the verb v is an ordered set of sentences s_i, such that $\forall s_i \; \exists s_k \in E_v \; and \; |k-i| <= \epsilon$, ϵ is a non-negative integer. All the verbs from the text T are divided into three groups. The first group contains rare verbs V_{unic}. The number of occurrences $|E_v|$ in the text for these verbs is below the border β: $|E_v| < \beta$. The second group includes common verbs V_{common}. These verbs get the largest values of $|E_v|$, and are distributed relatively evenly within the text. Typically, these are parts of collocations from scientific speech style, such as "ОСУЩЕСТВЛЯТЬ" ("TO CARRY OUT"), "ПРОИЗВОДИТЬ" ("TO MAKE"). The third group contains verbs-markers V_{marker}. Those verbs-markers are present in the

text in sufficient quantities and are unevenly distributed. These verbs can also be parts of collocations from scientific speech style.

Let each sentence s_k of the text T be assigned to some cluster c from a finite set of clusters C. For example, the set of clusters can be obtained by clustering on the base of the distribution of parts of speech along the text (described in detail in [7]). Clusters are obtained automatically, so their boundaries can be defined with an error margin. Let c_v^ϵ be a subset of textual neighborhood T_v^ϵ, which belonging to cluster c: $c_v^\epsilon = T_v^\epsilon \cap c$.

The verb v_m is marker of cluster c, if $|c_{v_m}^\epsilon|/|T_{v_m}^\epsilon| > \sigma$ and $\forall a \in C |a_{v_m}^\epsilon|/|T_{v_m}^\epsilon| \leq \sigma$. The values of ϵ and σ are parameters of marker detection algorithm and depend on the method of obtaining clusters C.

The text nest of verb-marker v_m is the set of verbs: $N_{v_m}^\mu = \{v | E_v \cap T_{v_m}^\mu \neq \emptyset\}$.

3 Results and Conclusion

The mathematical model described is realized in algorithms for automatic labeling of text fragments and for construction of text nests of verbs. In the software complex the sentence s_k is implemented as a parse tree of Dialing parser[1]. The table 1 presents the lists of verbs for the three clusters. The window size for clustering $L = 120$ sentences. The algorithm parameters values is $\epsilon = 7$ sentences and $\sigma = 0.9$.

Table 1. Verbs-markers for clusters

Cluster annotation (expert)	**Verbs-markers (automatic extraction)**
Cluster 1. Description of the research objects: introduction definitions and process of snow formation.	СЛУЖИТЬ, СМОТРЕТЬ, ЗАВИСЕТЬ, ЯВЛЯТЬСЯ, ОПРЕДЕЛЯТЬ, ИМЕТЬ, ПРОИСХОДИТЬ (TO SERVE, TO WATCH, TO DEPEND, TO BE, TO DEFINE, TO HAVE, TO HAPPEN)
Cluster 2. Chapter about calculations and laboratory processing of research results, different tables of classifications, fragments about parameters measurement.	ВЫЧИСЛЯТЬ, ВЫЧИСЛЯТЬСЯ, ЗАПИСЫВАТЬСЯ (TO CALCULATE, TO BE CALCULATE, TO REGISTER)
Cluster 3. Observation methodology: observation areas marking, equipment and recommendations.	СОСТОЯТЬ, ПРИНИМАТЬСЯ, ИСПОЛЬЗОВАТЬ, РЕКОМЕНДОВАТЬ, БЫТЬ (TO CONSIST, TO BE TAKEN, TO USE, TO RECOMMEND, TO BE as a link-verb)

Let's consider the example of a nest for marker "ВЫЧИСЛЯТЬ" ("TO CALCULATE") for $\mu = 7$ sentences:

[1] http://aot.ru/

НАПОМНИТЬ, ОТСУТСТВОВАТЬ, РАССЧИТЫВАТЬСЯ, ОКРУГЛЯТЬ, ЗАПАСТИ, ПРЕДСТАВЛЯТЬ, УЧИТЫВАТЬСЯ, ОКАЗАТЬСЯ, ПРОБИВАТЬСЯ, ПОЗВОЛЯТЬ, ВЫБИРАТЬ, ПОДСЧИТЫВАТЬ (TO REMIND, TO BE ABSENT, TO BE CALCULATED, TO ROUND, TO STORE, TO PRESENT, TO BE CONSIDERED, TO APPEAR, TO BREAK THROUGH, TO ALLOW, TO CHOOSE, TO COUNT)

The nest obtained shows that the set of verbs, which is located around the marker, belongs to the calculations and laboratory processing of research results for snow cover observations.

The algorithm of fragments labeling and nests construction has been checked using the test set containing 15 scientific texts of various authors and subjects. Selected verbs-markers are consistent with the expert annotation of the fragments content. Verbs-markers can be used for semantic labeling of automatically separated fragments, although some of them have no semantic value and are just stylistic features of a specific text. In the future we are planning to use verbs-markers to improve the accuracy of fragments boundaries determining. The nests of verbs received on the basis of the model presented will be used in algorithms of processes visualization using their text descriptions.

References

1. UzZaman, N., Allen, J.F.: Event and temporal expression extraction from raw text: First step towards a temporally aware system. Int. J. Semantic Computing **4**(4) (2010) 487–508
2. Wang, X., McKendrick, I., Barrett, I., Dix, I., French, T., Tsujii, J., Ananiadou, S.: Automatic extraction of angiogenesis bioprocess from text. Bioinformatics **27**(19) (2011) 2730–2737
3. Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F.: An Overview of Event Extraction from Text. In: Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011). Volume 779 of CEUR Workshop Proceedings., CEUR-WS.org (2011) 48–57
4. Johansson, R.: Natural Language Processing Methods for Automatic Illustration of Text. Licentiate Thesis. Department of Computer Science, Lund University, Lund, Sweden (2006)
5. Krayvanova, V., Kruychkova, E.: Automatic illustration of texts based on templates. In: Proceedings of All–Russian Conference "Knowledge - Ontology - Theory" (KONT-13) with international participatio. Volume 1. (2013) 235–240
6. E., M.: Automatic ontology learning from text document collection. In: Proceedings of Russian Conference on Digital Libraries. (2011) 293–298
7. Krayvanova, V., Kruychkova, E.: Application of automatic fragmentation for the semantic comparison of texts. In: 15th International conference SPECOM 2013 Proceedings, September 1-5. Lecture Notes In Artificial Intelligence, Springer (2013) 46–53
8. Lvov, A.: Linguistic analysis of the text and author recognition (2008)
9. Homutova, T.: Research text: integral analysis of lexis. Language and culture (4) (2010)
10. Rozental, D., Telenkova, M.: Glossary of linguistic terms. 2 edn. Prosveshenie, Moscow, Russia (1976)

Автоматический выбор глаголов-маркеров для задачи выделения описаний процессов в текстах на естественном языке

Варвара А. Крайванова

Алтайский государственный технический университет, Барнаул, Россия
krayvanova@yandex.ru

Аннотация В статье представлены промежуточные результаты исследования, конечной целью которого является разработка универсального алгоритма для автоматической визуализации диаграмм процессов по текстовым описанием этих процессов. Цель данного исследования — проверка возможности использования глаголов в качестве маркеров для семантической маркировки длинных фрагментов в научных текстах.

Ключевые слова: автоматическое фрагментирование текста, text mining, глаголы-маркеры, динамические параметры текста.

Visual Analytics in FCA-based Clustering

Yury Kashnitsky

Higher School of Economics, Moscow, Russia ykashnitsky@hse.ru

Abstract. Visual analytics is a subdomain of data analysis which combines both human and machine analytical abilities and is applied mostly in decision-making and data mining tasks. Triclustering, based on Formal Concept Analysis (FCA), was developed to detect groups of objects with similar properties under similar conditions. It is used in Social Network Analysis (SNA) and is a basis for certain types of recommender systems. The problem of triclustering algorithms is that they do not always produce meaningful clusters. This article describes a specific triclustering algorithm and a prototype of a visual analytics platform for working with obtained clusters. This tool is designed as a testing frameworkis and is intended to help an analyst to grasp the results of triclustering and recommender algorithms, and to make decisions on meaningfulness of certain triclusters and recommendations.

Keywords: visual analytics, formal concept analysis, triclustering, social network analysis.

1 Introduction

Classical Formal Concept Analysis (FCA) deals with data which describe a relationship between a set of objects and a set of attributes and provides methods to derive a concept hierarchy or formal ontology in them [1]. FCA is a powerful tool for revealing dependencies in data and is commonly applied to data mining (in particular, text mining), machine learning, knowledge management, semantic webs, software development, and biology.

As a natural extension of FCA, Triadic Concept Analysis (TCA) manages triadic data in a form of objects, their attributes, and conditions under which these objects have certain attributes [2]. A common example is a social network analysis with a context including users (objects), events they take part in (attributes) and interests (which might be regarded as conditions under which a user participates in a certain event).

As the task of finding all concepts or triconcepts is computationally challenging, certain relaxations of these terms have been introduced: biclusters [3] and triclusters [4]. Here we address triclusters, i.e. combinations of sets of objects, their attributes, and conditions where not every object must have each attribute. Triclustering provides an output in the form of object clusters with similar attributes under similar conditions. Therefore, it is applied to mining

users with common interests, applicants with similar competences or books labelled by close tags [5], [6]. Triclustering is also a basis for a certain type of recommender systems [7], [8].

Visual analytics is an increasingly popular branch of Computer Science which combines both human and computer qualities to solve a range of problems that might lay beyond the power of man or machine separately. Actually, it is a subdomain of data analysis focusing on decision-making through data preprocessing, data mining and interactive user interfaces. For instance, Siemens PLM software allows developers to collect, process, visualize report data in the 3D engineering environment, and make real-time decisions in the process of developing new vehicles. The same method is used in situational and decision-making centres, in nuclear power energetics, and in crime investigations.

In this paper, we explore these topics and describe a framework which uses visual analytics to solve some problems in FCA.

2 Visual analytics

2.1 Definition and specificity

Generalizing and selecting crucial aspects of various definitions of visual analytics [9], [10], here we propose the following one:

Visual analytics is a subdomain of data analysis focusing on analytical reasoning on the basis of interactive user interfaces in process of data mining, data preprocessing, knowledge representation, discovering dependencies, and decision-making.

Let us further consider core peculiarities of visual analytics and the tasks it is designed to solve: [11]

1. Visual analytics usually deals with complicated problems with big amounts of data requiring both human and machine resources.
2. The final goal of visual analysis is to enable users to obtain deep insight in problems to be solved which might include processing of large amounts of data from various sources. For this purpose visual analytics combines both human and technological resources. On one hand, data mining and statistics are the driving force of any automatic data analysis. On the other hand, human brain's aptitude for information perception and discovering dependencies in data complies to machine techniques and thus turns visual analytics into a promising sphere for further development.
3. In its development, visual analytics fosters in its turn the development of data mining, data representation and visualization, and analytical reporting.
4. Visual analytics also deals with human cognition, information perception, Computer Science, interactive and graphical design.
5. Visual analytics combines methods of information visualization and graphical data representation where visualization fosters human perception by the following means:
   (a) Enlarging data resources capacities makes user memorize less

(b) Reducing search, such as by representing a large amount of data in small space
(c) Enhancing recognition of patterns, such as when information is organized in space by its time relationships
(d) Supporting easy relationship inference
(e) Monitoring large amounts of potential events
(f) Providing techniques for dynamic data monitoring

2.2 Siemems

Siemens uses visual analytics techniques in its product lifecycle management (PLM) software enabling developers to collect, process, visualize report data in the 3D engineering environment, and make real-time decisions in the process of developing new vehicles. [1]

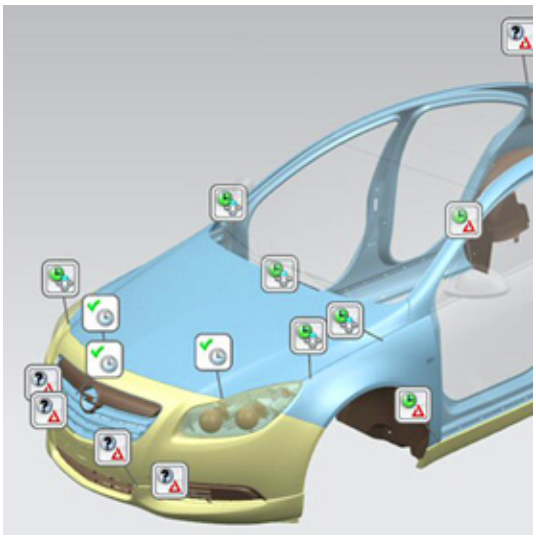


Fig. 1. One of development stages with Siemens PLM Software

The crucial point is that this system allows real-time visual interaction. This speeds up the processes of testing production for meeting given criteria, and eliminating product quality problems.

2.3 Supernova modelling

A highly powerful implementation of visual analytics paradigm was fulfilled by astrophysicists in Terascale Supernova Initiative (TSI) project. [2] The goal of

[1] http://www.plm.automation.siemens.com
[2] science.energy.gov/∼/media/ascr/ascac/pdf/meetings/mar03/Mezzacappa.pdf

the project is to give scientists from various fields access to powerful computation resources in order to produce knowledge in the sphere of fundamental science. In particular, the question of supernova birth was studied which encompassed 3D turbulence, gravitation and magnetic field modelling. The scale of the investigation was impressive - the modelling resulted in terabytes of data. The analysis of such amount of data lays beyond human power but combining human and machine capabilities allowed to make some inferences from all the bulk of data.

3 Formal Concept Analysis and OA-biclustering

3.1 Main definitions

A *formal context* in FCA is a triple $K = (G, M, I)$ where G is a set of objects, M is a set of attributes, and the binary relation $I \subseteq G \times M$ shows which object possesses which attribute. gIm denotes that object g has attribute m. For subsets of objects and attributes $A \subseteq G$ and $B \subseteq M$ *Galois operators* are defined as follows:

$$A' = \{m \in M \mid gIm \ \forall g \in A\},$$
$$B' = \{g \in G \mid gIm \ \forall m \in B\}.$$

A pair (A, B) such that $A \subset G, B \subset M, A' = B$ and $B' = A$, is called a *formal concept* of a context K. The sets A and B are closed and called the *extent* and the *intent* of a formal concept (A, B) respectively. For the set of objects A the set of their common attributes A' describes the similarity of objects of the set A and the closed set A'' is a cluster of similar objects (with the set of common attributes A').

The number of formal concepts of a context $K = (G, M, I)$ can be quite large ($2^{min\{|G|,|M|\}}$ in the worst case), and the problem of computing this number is #P-complete [12]. There exist some ways to reduce the number of formal concepts, for instance, choosing concepts by stability, index or extent size [13].

An alternative way is to make a relaxation of the definition of a formal concept. One of them is an OA-bicluster [3].
If $(g, m) \in I$, then (m', g') is called an *object-attribute bicluster* with the *density*

$$\rho(m', g') = \frac{|I \bigcap (m' \times g')|}{(|m'||g'|)}.$$

Bicluster density represents the percent of object-attribute pairs from the initial context in a certain bicluster.

Here are the main properties of OA-biclusters:

1. For any bicluster $(A, B) \subseteq 2^G \times 2^M$ it is true that $0 \leq \rho(A, B) \leq 1$,
2. An OA-bicluster (m', g') is a formal concept if $\rho = 1$,
3. If (m', g') is a bicluster, then $(g'', g') \leq (m', m'')$.

A bicluster (A, B) is called *dense* if its density is greater than a predefined minimum threshold, i.e. $\rho((A, B)) \geq \rho_{min}$. The above mentioned properties show that OA-biclusters differ from formal concepts since unit density is not required. Below follows an illustrative example for triconcepts and triclusters.

4 Triadic FCA and OAC-triclustering

As a solution for three-way data in FCA, Triadic Concept Analysis (TCA) was introduced [2].

A triadic context $K = (G, M, B, I)$ consists of sets G (objects), M (attributes), B (conditions), and ternary relation $I \subseteq G \times M \times B$. An incidence $(g, m, b) \in I$ shows that the object g has the attribute m under condition b.

We denote a triadic context by (X_1, X_2, X_3, I). A triadic context $K = (X_1, X_2, X_3, I)$ gives rise to the following dyadic contexts:

$$K^{(1)} = (X_1, X_2 \times X_3, I^{(1)}),$$
$$K^{(2)} = (X_2, X_3 \times X_1, I^{(2)}),$$
$$K^{(3)} = (X_3, X_1 \times X_2, I^{(3)}),$$

where $gI^{(1)}(m, b) \Leftrightarrow mI^{(1)}(g, b) \Leftrightarrow bI^{(1)}(g, m) \Leftrightarrow (g, m, b) \in I$.

The derivation operators (or prime operators) induced by $K^{(i)}$ are denoted by $(.)^{(i)}$. For each induced dyadic context we have two kinds of derivation operators. That is, for $\{i, j, k\} = \{1, 2, 3\}$ with $j < k$ and for $Z \subseteq X_i$ and $W \subseteq X_j \times X_k$, the (i)-derivation operators are defined by:

$Z \to Z^{(i)} = \{(x_j, x_k) \in X_j \times X_k \mid x_i, x_j, x_k$ are related by I for all $x_i \in Z\}$,

$W \to W^{(i)} = \{x_i \in X_i \mid x_i, x_j, x_k$ are related by I for all $(x_j, x_k) \in W\}$

A *triadic concept* of a triadic context $K = (G, M, B, I)$ is a triple (A_1, A_2, A_3) of $A_1 \subseteq X_1$, $A_2 \subseteq X_2$, $A_3 \subseteq X_3$ such that for every $\{i, j, k\} = \{1, 2, 3\}$ with $j < k$ we have $A_i^{(i)} = (A_j \times A_k)$.

A_1, A_2 and A_3 are called the *extent*, the *intent* and the *modus* of (A_1, A_2, A_3).

A set $T = ((m, b)', (g, b)', (g, m)')$ for a triple $(g, m, b) \in I$ is called an *OAC-tricluster* (or object-attribute-condition tricluster or just tricluster) based on prime operators. Here

$$(g, m)' = \{b \mid (g, m, b) \in I\},$$
$$(g, b)' = \{m \mid (g, m, b) \in I\},$$
$$(m, b)' = \{g \mid (g, m, b) \in I\}.$$

The *density* of a tricluster (A, B, C) of a triadic context $K = (G, M, B, I)$ is given by the fraction of all triples of I in the tricluster, that is
$\rho(A, B, C) = \frac{|I \bigcap A \times B \times C|}{|A||B||C|}$.

The tricluster $T = (A, B, C)$ is called *dense* if its density is greater than a predefined minimum threshold, i.e. $\rho(T) \geq \rho_{min}$. Just similarly to biclusters, triclusters have the following properties:

1. For every triconcept (A, B, C) of a triadic context $K = (G, M, B, I)$ with nonempty sets A, B and C we have $\rho(A, B, C) = 1$,
2. For every tricluster (A, B, C) of a triadic context $K = (G, M, B, I)$ with nonempty sets A, B and C we have $0 \leq \rho(A, B, C) \leq 1$.

4.1 Example

Let us consider a sample context $K = (U, I, S, Y)$, where U = {Ed, Leo, Max} is a set of users, I = {soccer, hockey} — their interests, S = {soccer.com, nhl.com, fifa.com, hockeycanada.ca} — sites they have added to bookmarks, $Y \subseteq U \times I \times S$ is a ternary relation between U, I, S which can be expressed by Table 1:

	i_1	i_2
u_1	x	x
u_2	x	x
u_3	x	x

	s_1	s_2	s_3	s_4
u_1	x	x	x	x
u_2	x	x	x	
u_3	x	x	x	x

	s_1	s_2	s_3	s_4
i_1	x		x	
i_2		x		x

Table 1. Sample context. Designations: u_1 - Ed, u_2 - Leo, u_3 - Max, i_1 - soccer, i_2 - hockey, s_1 - soccer.com, s_2 - nhl.com, s_3 - fifa.com, s_4 - hockeycanada.ca.

Here, generally, we have $|U||I||S| = 24$ triples to analyze. But actually, this number is reduced to 11, as there are lots of void triples present.

Actually, users Ed, Leo and Max share the same interests and almost the same sites (all the difference is that Leo has not bookmarked hockeycanada.ca). The idea of clustering here is presented by a tricluster $T = (\{u_1, u_2, u_3\}, \{i_1, i_2\}, \{s_1, s_2, s_3, s_4\})$ with density $\rho = 11/24 \cong 0.46$.
It is just one pattern to analyze instead of 11 in case of triples.

5 Implemented algorithms

The algorithms, described below, were implemented in Python 2.7.3 on a 2-processor machine (Core i3-370M, 2.4 HGz) with 3.87 GB RAM. One can find a description of testing procedure for these algorithms in [14] and [15].

5.1 OAC-prime triclustering algorithm

The hard core of the algorithm is quite simple: for all incidences $(g, m, b) \in I$ for a triadic context $K = (G, M, B, I)$ we build a tricluster $T = ((m, b)', (g, b)', (g, m)')$. If a tricluster is unique and its density exceeds a predefined minimum threshold then it is added to an array of triclusters. A pseudocode of algorithm for OAC-triclustering based on prime operators is presented below.

Algorithm 1 OAC-triclustering based on prime operators

Input: $K = (G, M, B)$ - tricontext,
ρ_{min} - density threshold
Output: $Tdic = \{X_1, X_2, X_3\}$ — a tricluster dictionary. $X_1 \subseteq G, X_2 \subseteq M, X_3 \subseteq B$

```
for (g, m, b) ∈ I do
    T = ((m, b)', (g, b)', (g, m)')
    HashKey = hash(T)
    if HashKey ∉ Tdic.keys() and ρ(T) ≥ ρ_min then
        Tdic[hashKey] = T
    end if
end for
```

5.2 Recommender algorithm based on triclustering

Algorithm 2 Recommender algorithm

Input: $K = (U, T, R, I)$ - tricontext, Tr - a set of triclusters
Output: Tag_{rec}, Res_{rec} - sets of recommended tags and resources

```
for u ∈ U do
    for i = 1,...,|Tr| do
        sim_u(Tr_i) = 1/2 ( |R_u ∩ R_Tr_i| / |R_u ∪ R_Tr_i| + |T_u ∩ T_Tr_i| / |T_u ∪ T_Tr_i| )
        Tr_best = argmax(sim_u(Tr_i))
        Tag_rec[i] = T_Tr_best \ T_u
        Res_rec[i] = R_Tr_best \ R_u
    end for
end for
```

The recommender algorithm applied to sets of a tricontext is analogous to the one described in [7]. It takes as an input a context of three sets (objects, attributes, conditions), and the set of triclusters obtained as a result of the OAC-prime triclustering algorithm. For each user among all triclusters the one most similar to triples with this user is selected. The similarity of triclusters and triples is defined by function $sim_u(Tr_i)$. The algorithm returns sets Tag_{rec}, Res_{rec} - tag and resource recommendations for all users.

6 The challenge and visual tricluster analysis framework

The challenge of the problem of triclustering (as of clustering on the whole) is to output meaningful, well-interpreted clusters. Actually, the term "meaningful" is not formally defined and is used by people to express their own subjective opinion on how well the task of clustering is solved, i.e. how similar the objects

in same clusters are, how distant - in different ones, how it corresponds to real world problems etc. Therefore, here an expert opinion might be useful, and a prototype of a visual analytics framework, described below, provides visual feedback for expert, and gives him ability to explore clusters in details.

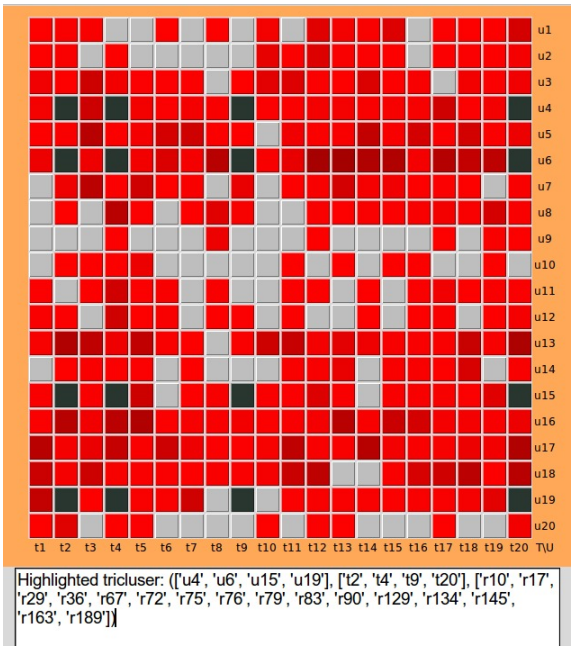


Fig. 2. Highlighting a largest tricluster for a user-tag pair $(u6, t4)$

In figure 2, we can see a map of triclusters produced by algorithm 1 for a context of 20 users, 20 tags, and 200 resources. The map is projected on the User-Tag plane. The more a certain user-tag pair is presented in triclusters the darker the corresponding square. A user-tag pair $(u6, t12)$, for instance, is included in 73 triclusters (a dark red square) while $(u5, t9)$ - just in 1 (a red square), and no triclusters have a pair $(u9, t10)$ (a grey one).

All triclusters including a certain user-tag pair can be listed by clicking on the "Triclusters" menu label. Similarly, triconcepts can be listed. One can also highlight the biggest tricluster with a certain user-tag pair or output all triclusters of the initial context ordered by density. Moreover, through the "Recommend attributes" context menu option an analyst can depict the results of recommender algorithm for a certain user (in this case, to show recommended tags).

The tool is intended to help an analyst to grasp the results of triclustering and recommender algorithms, and to make decisions on meaningfulness of certain triclusters and recommendations. The map helps the expert to quickly detect the concentrated regions (dark squares) and visualize dense triclusters including the

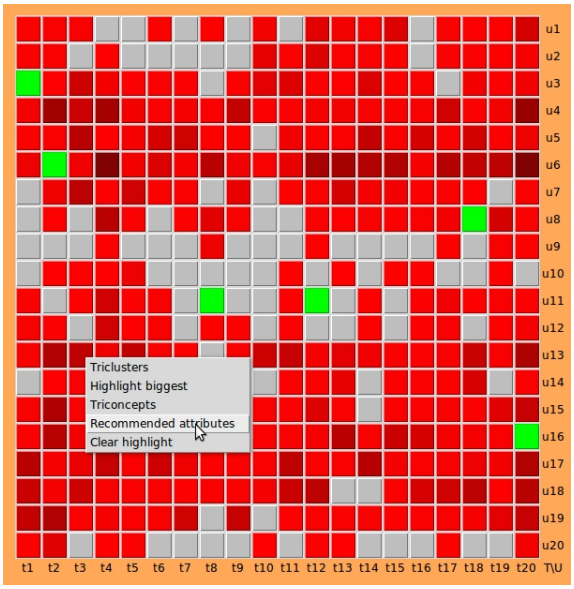


Fig. 3. Recommended tags for several users

corresponding triples. Further, it helps to make the decision whether the selected dense tricluster is meaningful or not, i.e. if it really combines similar users, tags, and resources.

7 Further work

There are several important issues to be regarded:

1. Limited human contribution: human contribution to triclustering in this visual analytics approach is limited and might only reach some hundreds of decisions on certain triclusters (less plausible, a thousand). Therefore, machine learning approach might help to learn the algorithm to classify meaningful clusters. The distance metric on triclusters should be carefully chosen.
2. Scalability: the issue of scalability is quite challenging in the described technique, and is to be solved. In current state, the application can support only contexts with one long dimension, for instance, a context of 20 users, 20 tags, and 400000 resources which can be projected onto a user-tag plane. One possible way to address the scalability issue is to perform preliminary clustering of objects, attributes, and conditions separately, and then choose representatives from each cluster.
3. Extending the idea of a human-machine approach to other problems in FCA or data mining, such as exploring implications and association rules in order to find meaningful ones.

8 Conclusion

Visual analytics, as one of the flourishing domains of data analysis, can be useful in mining objects with similar attributes under similar conditions in a context of social network data. A special algorithm was developed for uniting such objects, attributes, and conditions in triclusters. The program framework under development is intended to graphically display the results of this algorithm and to empower an analyst to decide on the meaningfulness of clusters and tags or resources recommendations for objects.

Acknowledgements The author would like to thank his colleagues from Higher School of Economics Sergei Kuznetsov and Dmitry Ignatov for their well-timed advice and support during this work. He also expresses gratitude to Stanislav Klimenko from Moscow Institute of Physics and Technology for consulting in visual analytics.

References

1. *Ganter, B., Wille, R.*: Formal concept analysis: Mathematical foundations. Springer, Berlin (1999)
2. *Lehmann F., Wille R.*: A triadic approach to formal concept analysis. Springer-Verlag, London (1995)
3. *Mirkin, B. G.*: Mathematical Classification and Clustering. Kluwer Academic Press, Dordrecht (1996)
4. *Ignatov, D. I., Kuznetsov, S. O., Poelmans, J., Zhukov, L. E.*: Can triconcepts become triclusters? International Journal of General Systems. 42, 572—593 (2013)
5. *Gnatyshak, D. V., Ignatov, D. I., Semenov, A., Poelmans, J.*: Analysing online social network data with biclustering and triclustering. In: Proceedings of the "Concept Discovery in Unstructured Data" conference, vol. 871, pp. 30—39. Katholieke Universiteit Leuven, Leuven (2012)
6. *Ignatov, D. I., Kuznetsov, S. O., Poelmans, J.*: Concept-Based Biclustering for Internet Advertisement. In: ICDM Workshops, pp. 123—130 (2012)
7. *Venjega, A. B., Gnatyshak, D. V., Ignatov, D. I., Konstantinov, A. V.*: Recommender system for perfumes and their tags based on triclustering. In: Proceedings of the "Intellectual data processing" conference, pp. 601—605. Torus Press, Moscow (in Russian) (2012)
8. *Ignatov, D. I., Poelmans, J., Zaharchuk, V.*: Recommender System Based on Algorithm of Bicluster Analysis RecBi. In: CEUR Workshop proceedings of the "Concept Discovery in Unstructured Data" conference, vol 757, pp. 122—126 (2011)
9. *Keim, D., Andrienko, G. et. al.*: Visual analytics: Definition, process, and challenges. In: Information Visualization, vol. 4950, pp. 154—175 (1999)
10. *Thomas, J., Cook, K.*: Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press, New York (2005)
11. *Kielman, J., Thomas, J.*: Special Issue: Foundations and Frontiers of Visual Analytics. In: Information Visualization, vol. 8, pp. 239—314 (2009)
12. *Kuznetsov, S. O.*: On Computing the Size of a Lattice and Related Decision Problems. Order, vol. 18, no. 4, pp. 313—321 (2001)

13. *Kuznetsov, S. O.*: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence. 49, 101—115 (2007)
14. *Gnatyshak, D. V., Ignatov, D. I., Kuznetsov, S. O.*: From Triadic FCA to Triclustering: Experimental Comparison of Some Triclustering Algorithms. In: CLA 2013 Proceedings, pp. 249—260. University of La Rochelle (2013)
15. *Kashnitsky, Y. S.*: Visual analytics for multidimensional data triclustering. Proceedings of MIPT, vol. 6, no. 2(22) (in Russian, to be published) (2014)

Визуальная аналитика в задаче трикластеризации, основанной на анализе формальных понятий

Юрий Кашницкий

НИУ ВШЭ, Москва, Россия
ykashnitsky@hse.ru

Аннотация Трикластеризация — это способ обнаружения объектов со схожими свойствами в контексте из трех множеств сущностей. Например, в задаче анализа данных социальных сетей, такими множествами могут быть пользователи, их интересы и события, в которых они принимают участие. Трикластеризация здесь может помочь найти группы пользователей с похожими интересами и, делать им рекомендации событий на основе этих интересов. В статье описывается конкретный алгоритм трикластеризации и прототип программной платформы для визуального анализа полученных трикластеров.

Ключевые слова: визуальная аналитика, анализ формальных понятий, трикластеризация, анализ социальных сетей.

Analysis System of Scientific Publications Based on the Ontology Approach

Viacheslav Lanin, Svetlana Strinuk

National Research University Higher School of Economics, Perm, Russian Federation
lanin@perm.ru, strinuk@mail.ru

Abstract. The article describes an approach to scientific publications repository creation based on ontology approach and corpus linguistics methods, processing of unstructured data (scientific papers) using GATE. Implementation of discussed methods is intended to decrease significantly labor intensity of information search and analysis, provide operational use of information in research.

Keywords: ontology, GATE, scientific publication

1 Introduction

The number of academic publications has been growing day by day. It can be explained by the fact that the Internet has made a lot of publications and e-libraries available (RINC, Springer, ACM etc). Complexity of academic papers search on the particular subject is increasing in this regard. To improve search quality and speed information resources have to be systemized and well arranged, a user has to be given convenient navigating and query facilities.

To solve the task of text data processing statistics (latent semantic search), graph and ontological methods are implemented. Every method listed above has its drawbacks. Latent semantic analysis does not take into account semantics. Graph methods cannot be applied to academic search due to absence of evident links among documents. Ontologies have a limited application owing to lack of ontologies building; moreover building indexes and its support are time consuming. Semistructured character of information and heterogeneity of its sources involve implementing tools and methods of artificial intelligence to sort out the tasks of text data processing (text mining, Semantic Web technology and agent technology).

2 Document ontology

To search, analyze and classify, catalogue and store information efficiently consolidating knowledge about their content and structure is needed.

Information about the following aspects of electronic documents is critical: document size (format), document type, document layout (document structure).

While creating ontological resource notions about all three aspects of representing information are included in the document. Each element is described by ontology. Notions from different aspects should be interconnected therefore adjacent electronic document ontology is created. There are many projects of development document ontology (for example, Dublin core [4], project ontologies «docOnto» [3], Document ontology SHOE [5], Document Ontology of Research Centre Linked Data DERI [9], Muninn project document ontology [9]), but each existing document ontology has its advantages and disadvantages for solving our tasks. So, we create own ontology specialized on academic paper description.

3 Academic paper description

In this research most popular academic paper structure was analyzed. Rules describe the article in the academic journal as "original research, which should faithfully reflect the content and results of the research". Hypothesis should be put forward and evidence should be provided to prove it. The article normally provides clear accurate findings.

According to these recommendations most typical elements of academic papers were identified (see Table 1). Each section has a particular function; its tasks are formalized and described in guidelines and numerous article writing handbooks provide substantial information on writing each section. Understanding functions, which each section has, makes further identification of key article elements and search automation. Article title, authors' names, affiliation are not worth processing as they are unique elements of the structure. The key words set describes field of research in terminology terms, this set is relatively verified sample of the most frequent terms. Ontology of scientific publication were described on OWL language.

4 System implementation

The demands to these systems were augmentability, support of amount of languages, possibility to work with thesauruses and other ontological resources. After analyses of existing systems GATE [2] (General Architecture for Text Engineering) was chosen. GATE is a set of Java tools for natural languages processing. This open code system suits operations of processing texts of any size. It is necessary to note that in linguistic resources, which are used while working with GATE there are three types of data: documents, corpuses and annotations.

The following tasks are solved through means of GATE: organization of annotated storage of articles, implementation of mechanism of key words automated highlighting, realization of mechanism of automated structure analysis of publications in undirected formats introduction, identifying key structure elements of the article, identifying relationships between articles. Obviously, functional capabilities of GATE are limited; to solve all the mentioned tasks own solutions through implementation of GATE API are planned.

Table 1. Essential article parts

Article part	Description
Abstract	Abstract contains important information about most important sections of the article. It does not provide references. Normally, objectives, methods, procedures and the main conclusions are described.
Introduction History Background	Introduction focuses on providing sufficient information about the field of research that is why this section usually has a lot of references. Introduction also contains objectives and tasks of the article. This section refers to general situation in the field of research. Introduction provides specification of the scale of research; rationale of choice of methods, preliminary results and conclusion.
Previous research Literature review	The main function of Previous research/Literature review – is to analyze published sources, which illustrate one way or another the problem the article refer to. Previous research/Literature review might be written as a general review or a review of literature for a particular period of time.
Present Approach Objectives Hypothesis Model Analysis Methodology	A key article sector, varying from article to article, Present Approach/Objectives/Hypothesis/Model/Analysis Methodology section describes uniqueness of the approach to problem solution and approach development. Hypothesis and interpretation methodology of data collected are presented in this section. It gives detailed method, methodology and procedure description.
Results Statistical Analysis	Results/Statistical Analysis section gives the summary of the results/data sometimes tables, diagrams and other visuals are added.
Theoretical Implications Summary Conclusion	Theoretical Implications/Summary/Conclusion section is usually a final section of the paper containing critical analysis and interpretation of results.
References	References section provides references to sources organized in accordance with editorial guidelines and instructions
Appendix	Important nonintegrated data are placed in Appendix Section.
Acknowledgements	Acknowledgements section is typically placed at the beginning or in the end of the article and is expression of gratitude to all who helped in research, writing the article etc.

The first stage of processing the corpus is creating aggregate document storage and filling it with articles. It is necessary to provide convenient and efficient support (storage and adding) of raw documents. Creating a separate catalogue to store documents with rubrics identified by experts is vital to implement GATE in future. Besides, Alfresco is implemented in Java language, which makes ontology processing

components integration with Semantic Web tools easier, as these tools are implemented on this language.

Linguistic markup is one of the key concepts of corpus linguistics. Linguistic markup identifies texts various parameters, allowing to achieve intelligent search in corpus. Text markup allows include metadata attribute to texts and their components. Basic markup is provided by GATE ready functions: tokenization and paragraphs, sentences and words markup on its basis; morpho-syntactic analysis (identifying the part of speech). More complicated markup (bibliography, the credits, etc.) may be realized by GATE tools improvement. The set of key words represents the paper in general and characterizes the work from the point of its relevance. Therefore characterizing the text via key words is critical for efficient academic search. Identifying key concepts cannot be executed through basic functions of GATE that is why additional module with application of GATE API is implemented in the system. To develop research prototype frequency approach is used due to its ease of use.

5 Conclusion

Implementation of discussed methods is intended to decrease significantly labour intensity of information search and analysis, provide operational use of information in research, and increase the amount of information from different sources available for processing. The basic mechanism of the system is knowledge oriented which allows providing integrated solutions to the tasks. Now it can be seen that the basis for creation of the intellectual system supporting research and providing efficient feedback is developed.

Acknowledgements. The reported study is supported by RFBR, research project №14-07-31273.

References

1. Bird S., Liberman. M. A Formal Framework for Linguistic Annotation. Technical Report MS-CIS-99-01, University of Pennsylvania, Philadelphia, PA, 1999.
2. Cunningham H., Maynard D., Bontcheva K. Text Processing with GATE. – Gateway Press CA, 2011.
3. CNXML/DocumentOntology http://mathweb.org/wiki/CNXML/DocumentOntology
4. Dublin Core Metadata Element Set, Version 1.1 http://dublincore.org/documents/dces/
5. Document Ontology (draft) http://www.cs.umd.edu/projects/plus/SHOE/onts/docmnt1.0.html
6. Grishman. R. TIPSTER Architecture Design Document Version 2.3. Technical report, DARPA, 1997. http://www.itl.nist.gov/div894/894.02/related_projects/tipster/.
7. Muninn Documents Ontology http://rdf.muninn-project.org/ontologies/documents.html
8. XML Languages http://cnx.org/help/authoring/xml
9. Varma P. Project Documents Ontology http://vocab.deri.ie/pdo.

Система анализа научных публикаций на основе онтологического подхода

Вячеслав В. Ланин, Светлана А. Стринюк
Национальный исследовательский университет «Высшая школа экономики»
vlanin@live.com,strinuk@mail.ru

Аннотация. В статье описывается подход к созданию хранилища научных публикаций с поддержкой семантического индексирования на основе онтологического подхода, методов компьютерной лингвистики и обработки неструктурированных данных. В качестве инструментальной среды для обработки текстов используется платформа GATE. Для анализа публикаций используются специально разработанные онтологические ресурсы, описывающие структуру публикаций и их формат. Также при обработке текстов используются словари ключевых слов и частотные характеристики текста. Реализация предлагаемого подхода позволит упростить поиск и анализ публикаций по заданной тематике, выявить связи между ними.

Ключевые слова. онтология, GATE, научная публикация.

Automatic Defect Recognition in Corrosion Logging Using Magnetic Imaging Defectoscopy Data

Rita Gaibadullina, Bulat Zagidullin, Vladimir Bochkarev

Kazan Federal University, Kazan, Russia
{rita.gaibadullina,bulatza}@gmail.com, vbochkarev@mail.ru

Abstract. The Magnetic Imaging Defectoscopy is designed for detection of corrosion zones in oil wells. Location of corrosion zones is a time-consuming process, during which some defects can be missed. Therefore this process shall be automated. This document describes an algorithm of automatic defect recognition based on maximum likelihood criterion and the use of wavelet threshold processing for noise reduction and pre-conditioning of experimental data.

Keywords: Magnetic Imaging Defectoscopy (MID), wavelet filtering, maximum likelihood criterion.

1 Introduction

The Magnetic Imaging Defectoscopy can be used to identify defects, corrosion intervals in oil wells. The tool generates an electromagnetic pulse and receives time-related response of tubing and casing walls. The attenuation rate of the response depends on the electromagnetic characteristics of the tube material and its thickness. Metal loss due to corrosion causes a faster decay than non-corroded metal.

The Magnetic Imaging Defectoscope (MID) contains of two sensors: the short and the long sensors. The short sensor is 120 mm in length designed to sense the tubing. The sensor generates a short pulse (50 ms) of low amplitude and magnetises basically the first barrier only, and then receives the response of 0.1 ms to 75 ms. Each decay of the short sensor consists of 42 points. The long sensor is 320 mm long; it generates pulses of greater amplitude and duration (250 ms) and takes the total response from both the first and second barriers (tubing and casing) within 275 ms. Each decay of the long sensor contains 51 points. Thus, the experimental data are presented by the 42 logs for the short sensor and 51 logs for the long sensor (Fig. 1).

Each log of the long or short sensors can be divided into the trend and drift components. The trend means a log component slowly varying with depth (can be found, for example, using a median filter). The drift means a component rapidly varying with depth, which shows deviation of real log from the trend [1].

$$A_{drift} = \frac{A(t) - A_{trend}(t)}{STD} \quad (1)$$

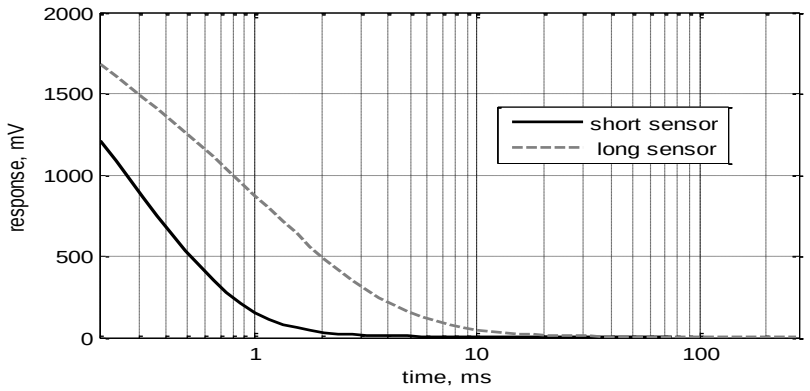


Fig. 1. Responses of the short and long MID sensors.

A DRIFT panel shall be built for visualisation of a drift components normalised to the standard deviation (STD). Generally, it is a three-dimensional graph, where the vertical axis shows the depths, the horizontal axis shows the decay time and the colour determines the signal amplitude (see Fig. 2). Gain in signal at depths of X835 ft and X820 ft corresponds to the tubing and casing collars, respectively. Gain in signal at depths of X855 ft also corresponds to the casing collars. Reduction in signal at the depth of X837 ft displays casing corrosion, which can be detected through the tubing.

Nowadays, corrosion zones are detected during well log analysis, i.e. their location is arbitrary. Moreover, the analysis of 6,000 – 9,000 ft wells consumes plenty of time, during which the defects can be missed. Therefore solution to this problem is automation of corrosion interval detection.

2 Automatic Recognition of Corrosion Intervals

2.1 Wavelet Filtering of DRIFT Data

Data are pre-filtered to remove the noise components, which could affect the performance of the recognition algorithm.

A two-dimensional wavelet decomposition is applied to DRIFT data. This wavelet decomposition is designed for processing of two-dimensional pictures with commensurable number of points in X- and Y-directions. In our case, the number of counts in the vertical axis (i.e. well depth) has an order of thousands that ten and hundred times greater than the number of values of horizontal axis (totally 42 and 51 time-related counts), therefore the two-dimensional wavelet decomposition is used first and then the one-dimensional wavelet decomposition in X-direction. The threshold value is calculated by two methods: Donoho and Birgé-Massart strategies [2, 3].

3 Algorithm of automatic corrosion recognition

An algorithm of automatic defect recognition includes two main steps:

1. Construction of binary maps according to DRIFT panels;
2. Making a decision on significant deviation on binary maps;

1. Construction of binary maps using data from DRIFT panels. Statistical DRIFT data $\xi_{t,d}$, are converted into binary maps by some THR threshold. It is necessary to take into account that increase of the signal corresponds to the presence of collar:

$$\eta_{t,d} = \begin{cases} 1 & \text{if} \quad \xi_{t,d} > THR \\ 0 & \text{if} \quad \xi_{t,d} \leq THR \end{cases}, \tag{2}$$

and decrease of the signal, on the contrary, corresponds to the presence of corrosion:

$$\eta_{t,d} = \begin{cases} 1 & \text{if} \quad \xi_{t,d} < -THR \\ 0 & \text{if} \quad \xi_{t,d} \geq -THR \end{cases}, \tag{3}$$

2. The automatic defect recognition process is based on the decision theory. There are two hypotheses: H_0 – the defect is absent and H_1 – the defect is present.

$$P(x = 1 \,/\, H_0) = \alpha, P(x = 0 \,/\, H_0) = 1 - \alpha, \tag{4}$$

$$P(x = 1 \,/\, H_1) = 1 - \beta, P(x = 0 \,/\, H_1) = \beta, \tag{5}$$

where α - error of first kind, β - error of second kind.
Each hypothesis has its likelihood function. A value of the likelihood logarithm is calculated for each hypothesis at each depth point.

The defect is absent:

$$l(0) = \sum_{t \in I} \eta_{t,d} \ln \alpha + \sum_{t \in I} \left(1 - \eta_{t,d}\right) \ln(1 - \alpha) \tag{6}$$

The defect is present:

$$l(1) = \left\{\sum_{t \in I} \eta_{t,d}\right\} \ln(1 - \beta) + \left\{\sum_{t \in I} \left(1 - \eta_{t,d}\right)\right\} \ln \beta. \tag{7}$$

Then the two-decision statistical hypothesis is verified by the maximum likelihood method:

$$l(1) - l(0) \underset{H0}{\overset{H1}{\gtrless}} C, \tag{8}$$

where $\underset{H0}{\overset{H1}{}}C = 0$ - the decision threshold by the maximum likelihood criterion [4].
Figure 2 illustrates an example of corrosion in the casing, as the defect appears on the LONG DRIFT panel. The algorithm correctly identified the presence of corrosion and referred it to the corrosion of the first barrier, which is the tubing.

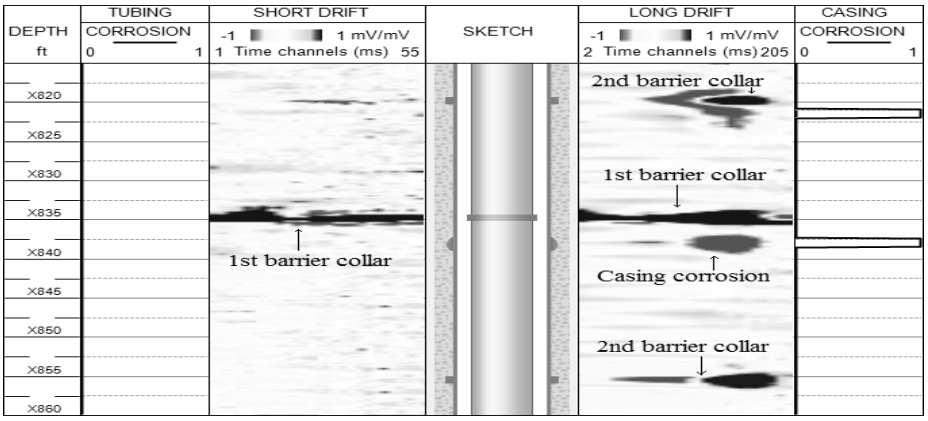


Fig. 2. Corrosion in the casing. Left to right: The DEPTH panel, TUBING CORROSION shows corrosion in the tubing, SHORT DRIFT is the drift panel for the short sensor, WELL SKETCH depicts well completion, LONG DRIFT is the drift panel for the long sensor, and CASING CORROSION shows corrosion in the casing.

In order to verify the algorithm, data from 11 wells were processed. Corrosions found during the well log analysis were compared with those processed by the algorithm. The following results were obtained: the automatic defect recognition algorithm accurately separates the defects of the 1st and 2nd barriers. When configuring the algorithm to search for small intervals of corrosion (metal loss less than 10%), lots of false defects are indicated, which complicates data processing. With such configuration, the algorithm detects 89% of the first barrier corrosions and 93% of the second barrier corrosions. Defects with metal loss less than 10% are not dangerous, unlike major defects with metal loss greater than 10%. The algorithm is designed to find major defects. When setting the appropriate algorithm parameters, all defects, including a small number of false defects, are detected. Thus, the automatic defect recognition allows quick identification of probable corrosion zones, on which the well log analyst should focus. This, in its turn, increases the speed and quality of data interpretation.

References

1. Arbuzov, A.A., Bochkarev, V.V., Bragin, A.M., Maslennikova, Yu.S., Zagidullin, B.A., Achkeev, A.A., Kirillov, R.S.: SPE 162054 Memory Magnetic Imaging Defectoscopy (2012).
2. Mallat, S.: A wavelet tour of signal processing. PP. 479-481. Mir Press, Moscow (2005).
3. Birgé, L., P. Massart. From model selection to adaptive estimation, in D. Pollard(ed), Festschrift for L. Le Cam, pp. 55–88, Springer (1997).
4. Lehmann, E.L.: Testing Statistical Hypotheses. P. 24. Nauka Press, Moscow (1979).

Автоматическое обнаружение дефектов и коррозии нефтяных скважин по данным магнитно-импульсного дефектоскопа

Рита Гайбадуллина, Булат Загидуллин, Владимир Бочкарев

Казанский (Приволжский) Федеральный Университет
{rita.gaibadullina,bulatza}@gmail.com, vbochkarev@mail.ru

Аннотация. Магнитно-импульсная дефектоскопия предназначена для выявления различных дефектов и интервалов коррозии. Высокочувствительные датчики, представляющие собой приёмно-возбуждающие катушки, позволяют анализировать отклик от окружающей среды в широком диапазоне времен. В настоящее время определение зон коррозии осуществляется интерпретатором, т.е. носит субъективный характер. Более того, анализ 2.5–3 км скважины (около 300 трубок НКТ и колонны) — это трудоёмкий процесс, в ходе которого часть дефектов может быть пропущена. Для исключения такого рода ошибок необходимо автоматизировать процесс поиска интервалов коррозии. В работе предложен алгоритм автоматического распознавания дефектов, позволяющий разделить типичные и нетипичные отклики. Так же рассмотрено применение пороговой вейвлет-обработки для подавления шумов и предварительной подготовки экспериментальных данных к дальнейшей обработке.

Ключевые слова: магнитно – импульсная дефектоскопия, вейвлет-фильтрация, критерий максимального правдоподобия.

Automated Generation of Assessment Test Items from Text: Some Quality Aspects

Andrey Kurtasov

Vologda State University, Vologda, Russia
akurtasov@gmail.com

Abstract This paper overviews the problem of automated generation of assessment test items from natural-language text. In a previously published article, an experimental system aimed at generating fill-in-the-blank test items from Russian text was decribed. In this paper, some aspects of the system's quality are analyzed. Main directions for future work are defined, including evaluation of the system and development of methods for filtering text fragments and selecting words to blank out.

Key words: educational assessment, natural language processing, Russian language, test item generation, question generation.

1 Introduction

The teaching process of today widely uses electronic text resources that were not originally intended for use as teaching aids. This is especially true for subjects that deal with rapidly developing domains such as information technology. Teaching these subjects may benefit from use of various articles and technical papers, which do not contain test questions or exercises, as opposed to textbooks. Developing the exercises is a complex task that may require a teacher to spend a significant amount of time on. A promising way to facilitate this task is automated generation of test items from text with the help of Natural Language Processing (NLP).

The general idea is to extract fragments from the source text document and to transform them into questions or test items. This idea has been studied by several researchers, and is commonly considered difficult to implement. For instance, Heilman [1] has discovered numerous challenges in question generation from text. These include linguistic challenges (lexical, syntactic, discourse-related) as well as various challenges related to the application of question generation tools in classrooms (usability, human-computer interaction issues).

Previously, we have described an experimental system for generating fill-in-the-blank test items from Russian-language text, which is designed for use with the e-learning platform Moodle[1] [2]. We have showed that the automated generation of test items is not accomplished easily, but can yield some useful results. In this paper, we are going to review the quality aspects of the approach being studied and consider ways to improve it.

[1] https://moodle.org/

2 Test Item Generation: Approach and Quality Aspects

We present the approach as the sequential application of text processors that perform the following operations on the document:

1. Text preprocessing — to convert a raw text file into a well-defined sequence of linguistically meaningful units (as defined in [3]), or segments
2. Segment filtering — to filter the set of segments so that it contains the most salient segments
3. Test item generation — to transform the text segments into test items

Let us consider each of the operations from the quality perspective.

2.1 Text Preprocessing

This operation consists of two stages: document triage and text segmentation. Document triage is the process of converting a digital file into a well-defined text document. It involves such actions as character encoding identification and text sectioning (identificating the actual content within a file while discarding headers, links, and formatting features). This stage is solely technical and easy to accomplish with available software tools. However, it could crucially affect the results (e.g. improper encoding detection would make the Russian text unreadable), and should be a significant concern to the software developers.

Text segmentation is performed to acquire segments from which to produce test items. Previously we referred to this stage as sentence splitting, because we use sentences as basic segments, while considering a sentence to be a semantically complete portion of text. At first sight, a sentence is a sequence of characters that ends with ".", "!", or "?", but in practice one should keep in mind that these characters can also be used inside one sentence [4]. Today's NLP tools perform sentence splitting with fairly high precision. In preliminary experiments we used a tokenization module provided by the AOT toolkit[2], which recognizes common Russian abbreviations with periods, such as "г." (year), "гг." (years), "и т. д." (etc.), "т. е." (i.e.), "т. н." (so called), as well as special text features such as bulleted lists, sentences enclosed in quotation marks or parentheses, and URLs. The experiments have shown that this step does not introduce a significant number of errors in the resulting test items.

In some cases, it may be reasonable to include more than one sentence in a segment (when multiple sentences are used to express one significant thought). While automatic detection of such sentence groups is a complex semantics-related task, we assume that the user should be given an ability to see the context of the processed sentence at the test item generation step. This ability would allow the user to expand the segment if needed, and should be considered for implementation during the user interface design of the generating software.

[2] `http://www.aot.ru/`

2.2 Segment Filtering

It is obvious that not every text sentence is appropriate for test item generation. We assume that proper filtering of acquired sentences could have a convincing impact on the quality of the resulting test items set, and we propose using extractive text summarization to filter out the unneccessary text portions. In NLP, different methods for scoring sentences by importance are applied (usually in combination) [5]: sentence length cut-off (short sentences are excluded), use of cue phrases (inclusion of sentences with phrases such as "in conclusion"), sentence position in a document/paragraph, occurrence of frequent terms (based on TF-IDF term weighting), and occurrence of title words.

We are planning to leverage an existing summarization toolkit and attempt taming it for our task. For example, MEAD[3] is claimed to be modifiable to support languages other than English. Similarly to the text segmentation, it would be reasonable to show the highest-scoring sentences inline, so that the user could see the discarded portions and use them if they appear to be useful.

The performance of this step is to be evaluated experimentally. We are planning to compare the summarization output with the selection made by human experts and calculate such metrics as precision and recall (commonly used in informational retrieval).

2.3 Test Item Generation

As a starting point of the research, we generate fill-in-the-blank test items ("cloze questions"). To produce a cloze question, we take a sentence and replace some of the words in the sentence with blanks. For additional clarity, we add a hint into the question, explaining what kind of answer is expected. Below is an example:

Source: *В отличие от перцептронов рефлекторный алгоритм напрямую рассчитывает адекватную входным воздействиям реакцию интеллектуальной системы.*
Result: *В отличие от перцептронов ..... (какой?) алгоритм напрямую рассчитывает адекватную входным воздействиям реакцию интеллектуальной системы.*

Or, in English:

Source: *In contrast to perceptrons, the reflective algorithm directly calculates the reaction of the intelligent system with respect to input actions.*
Result: *In contrast to perceptrons, the ..... (what?) algorithm directly calculates the reaction of the intelligent system with respect to input actions.*

The system recognized an adjective ("рефлекторный" — "reflective"), replaced it with a blank, and inserted a hint in parentheses: "какой?" ("what?"). Also, the current system is able to add appropriate hints for acronyms, numbers, definitions, sentence subjects, adverbials (more examples were shown in [2]).

[3] `http://www.summarization.com/mead/`

The main problem here is to determine which words should be blanked out to produce a useful question. A good approach could be finding the sentence's focus (in the sense of information structure), which is difficult to do with the state-of-the-art NLP tools. Another idea is based on the assumption that it is more useful to blank out special terms than common words. We could match words of the sentence against either a pre-existing domain-specific bag of words or a bag of words acquired through terminology extraction from the processed text, and blank out the matches.

Another issue, which arises at this step, is that the processed sentences may contain anaphora. Without an implementation of automatic anaphora resolution, the user could resolve the anaphora manually (e.g. to replace pronouns with corresponding nouns) using the in-context display of the processed sentence.

While cloze items are fairly easy to produce from sentences, fill-in-the-blank is a trivial style of test. This concern could be addressed by considering the two ideas: generation of interrogative sentences (it would require text simplification and word reordering [1]) and generation of distracting answers for multiple-choice tests (a possible solution is described in [6]).

3 Conclusion and Future Work

Based on the preceding research, we have considered several quality aspects of the automated generation of assessment test items from natural-language text. We have discovered the following directions for quality improvement of our system:

1. The user interface should display the context of the text excerpt being processed in a user-friendly way for efficient human-computer interaction.
2. We will leverage a summarization toolkit for segment filtering and evaluate it experimentally.
3. Other directions include anaphora resolution, interrogative sentence generation, and distractor generation for multiple-choice tests.

References

1. *Heilman, M.* Automatic Factual Question Generation from Text. Ph.D. Dissertation. — Carnegie Mellon University, Pittsburgh, USA, 2011. 195 p.
2. *Kurtasov, A.* A System for Generating Cloze Test Items from Russian-Language Text / In Proceedings of the Student Research Workshop associated with The 9th International Conference RANLP 2013. P. 107–112. — Hissar, Bulgaria, 2013.
3. *Indurkhya, N.; Damerau, F. J. (eds).* Handbook of Natural Language Processing (Second Edition). — Chapman and Hall/CRC, 2010. 704 p.
4. *Grefenstette, G.; Tapanainen, P.* What is a Word, what is a Sentence? Problems of Tokenisation / In Proceedings of The 3rd Conference on Computational Lexicography and Text Research. — Budapest, Hungary, 1994.
5. *Hynek, J.; Ježek, K.* Practical approach to automatic text summarization. / In Proceedings of the ELPUB 2003 conference. — Guimaraes, Portugal, 2003.
6. *Mitkov, R., Ha, L., Karamanis, N.* A computer-aided environment for generating multiple-choice test items // Natural Language Engineering. 2006. 12(2). P. 1–18.

Автоматизированная генерация тестовых заданий для проверки знаний: некоторые аспекты качества

Андрей Куртасов

Вологодский государственный университет, Вологда, Россия
akurtasov@gmail.com

Аннотация В работе приведен обзор задачи автоматизированной генерации тестовых заданий для проверки знаний из текста на естественном языке. В ранее опубликованной статье была описана экспериментальная система для генерации заданий на заполнение пропусков из русскоязычного текста. В данной работе проанализированы некоторые аспекты качества работы системы. Определены основные направления для дальнейшей работы, включая оценку системы и разработку методов фильтрации текстовых фрагментов и выбора слов для замены на пропуски.

Ключевые слова: оценка знаний в образовании, автоматическая обработка текста, генерация тестовых заданий, генерация вопросов.

GPS Navigation Algorithm Based on OSM Data

Daniel Khachay

Ural Federal University, Yekaterinburg, Russia
daniil.khachay@gmail.com

Abstract. A new pedestrian GPS navigator providing the shortest-cost safest-crossing route on the basis of Open Street Map (OSM) cartographic data is proposed. Also, Java implementation and use case example are discussed.

Keywords: GPS-navigation application, OSM data retrieval, shortest-path search algorithm

1 Introduction

Satellite navigation algorithms are used everywhere in modern life. Every computing system, even a smartphone, is equipped with some kind of navigation application (at least, Google Maps). Such an application is able to build a route from one point to another, show it on a map, etc. Each navigation application can be proprietary or open-source. Among wide variety of open-source projects, Open Street Map (OSM) project seems to be the most interesting. I've decided to study this format in more details. I know that the best way to understand a new technology better is to apply it for something useful. I'm fond of walking the streets of my city. So I decided to develop a simple pedestrian navigator based on OSM data.

2 Problem Statement and Related Works

Each navigation application contains implementation of some routing algorithm as it's main building block.

General routing problem has the following setting. Input: starting and target location points given by their GPS coordinates and topographic map of the search area, defining restrictions over feasible routes. The goal is to determine the optimal (shortest w.r.t some predefined metric) route.

For instance, a car driver navigator constructs a minimum trip-time route subject to given road map and traffic constraints.

Traditional approach to mathematical solution of the above problem consists of two stages. On the first stage, the initial problem is reduced to well-known Shortest Path Problem (SPP), which is defined on the appropriate weighted graph. On the second stage, SPP is solved by one of classical combinatorial optimization algorithms: Dijkstra [1] or it's heuristic extension - the A^* [2].

Therefore, any navigation application differs from another only by the following features:

(i) setting of the initial problem (car driver navigation, bicycle navigation, etc.),
(ii) method of the reduction to SPP,
(iii) format of cartographic data.

Functionality (hover for description)	OpenRoute-Service [1]	YOURS [2]	CycleStreets [3]	Cloudmade	Routino [4]	BBBike @ World [5]	MapQuest [6]	OSRM [7]	TripGo [8]	BRouter	OpenTrip-Planner [9]	HoofMarker	GraphHopper [10]
Coverage	Europe only	Global	UK only	Global	UK only	Selected Cities	Global	Global	Selected Cities	Global	Global	Germany	Global
Modes of transportation													
Car (fastest)	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Experimental	No	No	Yes
Car (shortest)	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Experimental	No	No	No
Bicycle (shortest)	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	No	No
Bicycle (fastest)	Partial	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Bicycle (safest)	Yes	No	No	No	Yes[1]	Yes	No	No	No	Yes	Yes	No	No
Bicycle (quietest)	Safest	Partial[2]	Yes	No	Yes[1]	No	No	Yes	No	Yes	No	No	No
Pedestrian	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Shortest	Yes	Yes	Yes
Pedestrian (safest crossings)		No	No		No	No	No	No	No	No	No	No	No

Fig. 1. The comparison matrix taken from the OSM project's official site [6]

.

Nowadays, there are many open-source applications based on OSM data format. Among them, the following applications seem to be the most popular:

a **Open Source Routing Machine** is a nice online routing application. It's seems to be [3] mostly fast and precise car driver navigator ever. Unfortunately, it have no standalone version and provides no services for bicyclers and pedestrians.
b **CycleStreets** is a mobile routing application for iOS and Android platforms [4]. The iOS version of the application is seems to be the most valuable for bicyclers. However, it operates only in United Kingdom and has no services for pedestrians like the previous application.
c **GraphHopper** is Java implemented cross-platform multi-service routing application [5], which seems to be the most interesting. The application provides simultaneously a car driver, a bicycler and a pedestrian navigation services. But the pedestrian navigator provides no a safest-crossings routing.

As can be seen from Fig. 1 there is no OSM-based routing application providing full service to pedestrians (no safest-crossings support). In this paper we describe our OSM-based Java implemented standalone application, providing this type of service.

Let us recall some basics of the OSM format structure. First of all, OSM file [7] is just a special type of an XML document and contains hierarchical collection of elements. Some of these elements may have attributes and additional data. Basic elements of any OSM file are called *nodes* and *ways*. A *node* is just a model of some location point, defined by geographic coordinates (latitude and longitude). In the OSM format, each spacial topographical object is presented

```
<way changeset="7806002" id="39240803" timestamp="2011-04-08T17:25:58Z"
uid="133332" user="AMDmi3" version="5" visible="true">
  <nd ref="470133843"/>
  <nd ref="804028626"/>
  <nd ref="470133865"/>
  <tag k="highway" v="pedestrian"/>
  <tag k="name" v="Tolmacheva st."/>
</way>
```

Fig. 2. An example of the pedestrian way description in OSM format

by some *way* element containing a collection of *nodes* and accompanied by informational elements. Each informational child element can be considered as a (key, value)-pair, presenting some feature of the containing *way* element (see Fig. 2).

3 Results

3.1 Reduction to the Shortest Path Problem in Weighted Graph

As we've seen above, in the OSM format, every topographical object (street, bicycle path, footpath, building, etc.) is presented by a piecewise linear way consisting of nodes. Here's an another reason for usage exactly the OSM format for the construction of the graph corresponding to the current navigation problem.

During the reduction to SPP, these nodes are just taken as vertexes of the constructed graph. Further, we assume two vertexes to be adjacent if they correspond to neighboring nodes of the same way on the map.

Because of our intention to develop the pedestrian (safest crossing) navigation application, we should consider only such ways, that describes special types of roads, among them footways, sidewalks, pedestrian crossings, etc. During the graph construction we use only these ways. When the weighted graph is constructed, we apply an heuristic extension of well-known Dijkstra algorithm - the A* algorithm [2] to construct a minimum cost path.

3.2 Java Implementation

We implement the mentioned above application in object-oriented Java application[8]. Let's describe the main classes. Main - the main class that runs an application. Footway - the "footpath" class, it contains an array of nodes from OSM file. AStar - the class that implementing an A* search algorithm. MapReader - this class is intended for parsing OSM file. RouteWriter - this class appends the calculated route (in osm-file) in XML-format so the augmented map could be visualized. WeightedPoint - this class determines the coordinates of nodes.

3.3 Example of Application Usage

Suppose, we are asked to construct a shortest pedestrian path from the main building of Institute of Mathematics and Computer Science (IMCS) UrFU to the

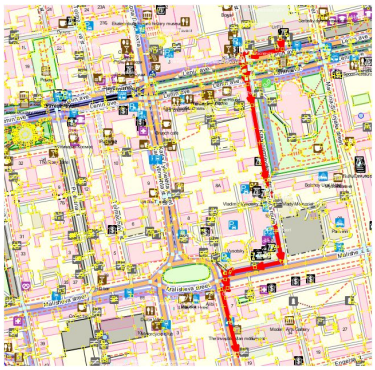


Fig. 3. The shortest pedestrian (safest crossings) path from IMCS UrFU to Belinsky public library is found and red-highlighted

main entrance of Belinsky public library. First we need to know GPS-coordinates of both way-points. Second we need a special OSM file that contains required piece of map. Then using the AStar class, our application constructing a shortest path and using a RouteWriter class, application write down an XML file with the shortest pedestrian route. To get the constructed route, we can open this updated XML file by any text-reading application. In our case (see Fig.3), we use JOSM (Java Open Street Map, Java implemented OSM editor) for graphical visualization.

We conduct a specific numerical experiment consisting of constructing of 100 routes for independently chosen random location points on the Ekaterinburg city map. Expected relative value of graph construction run-time is equal to 97.8% of total run-time within standard deviation of 0.3%.

4 Conclusion

A new type of OSM-based routing application for constructing shortest-cost safest-crossing pedestrian paths is proposed. The application is Java-implemented and can run on every Java-compatible platform. Run-time of the application can be significantly reduced by leveraging some of standard cashing techniques for a previously constructed graph.

References

1. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. Introduction to Algorithms (3-rd edition). MIT press. 2009.
2. Judea Pearl. Heuristics: Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley. 1984.
3. Open Sourse Routing Machine application page. http://map.project-osrm.org/

4. CycleStreets application page. http://www.cyclestreets.net
5. GraphHopper application page. http://graphhopper.com
6. OSM-based routing applications comparison matrix. http://wiki.openstreetmap.org/wiki/Routing/online_routers
7. Jonathan Bennett. OpenStreetMap. Packt Publishing. Birmingham. 2010.
8. Daniel Y. Liang. Introduction to Java programming: comprehensive version. Pearson Education, Inc. 2007.

Алгоритм GPS навигации по данным OSM

Даниил Хачай

Уральский федеральный университет, Екатеринбург, Россия
daniil.khachay@gmail.com

Аннотация Предлагается новое приложение — GPS-навигатор, находящее кратчайший маршрут с учетом правил дорожного движения по данным OpenStreetMap (OSM). Обсуждаются программная реализация на языке Java и пример использования.

Ключевые слова: GPS-навигатор, обработка OSM данных, алгоритм кратчайшего пути.

System of Ontologies for Data Processing Applications Based on Implementation of Data Mining Techniques

Alexander Vodyaho[1], Nataly Zhukova[2]

[1]Saint-Petersburg State Electrotechnical University, Saint Petersburg, Russia
[2]Saint-Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Saint Petersburg, Russia
{aivodyaho,nazhukova}@mail.ru

Abstract. The paper describes a system of ontologies developed for the applications oriented on solving problems of situations recognition and assessment based on results of data processing and analyses. Main attention is focused on the problems of processing measurements of various objects parameters represented in a form of time series. The considered applications process data using knowledge extracted from historical data with the help of Data Mining techniques. Such applications are highly knowledge centric and their core element is knowledge base that is represented as a system of ontologies. The proposed system of ontologies is a set of upper level ontologies for which techniques of adaptation for solving applied tasks for one or several related subject domains are developed.

Keywords: knowledge representation, data analyses, data fusion, measurements processing, situation recognition and assessment.

1 Introduction

Nowadays multiple problems in various subject domains are required to be solved at the level of situations [1, 2]. Results of solving problems at this level are much easier interpretable by an end user than results represented at lower levels of information generalization. Solving problems at the level of situations assumes solving such problems as recognition of situations, formal description of situations, analyses of situations, their estimation, assessment, prediction and awareness. Main sources of information about situations are results of measurements received from different types of instruments that measure parameters of technical and / or environmental objects. Real systems have to process huge volume of information including bad quality information. The majority of real life problems require that measurements are processed in real time or in the mode close to real time. It considerably increases the complexity of the problems. The problems can be solved with the desired quality and in limited time only using knowledge-oriented technologies. These intelligent technologies are based on application of data mining algorithms along with other means of artificial intelligence such as expert systems and inference machines. A set of basic solutions for

developing intelligent technologies for measurements processing (IMPT) and examples of their implementation are proposed in [3, 4, 5, 6].
The intelligent measurements processing technologies are described in general form using web ontology language (OWL). When new measurements are received an appropriate technology is selected and detailed using an a priori defined set of production rules. The rules are two part structures that use first order logic for reasoning over knowledge representation [7]. The detailed technologies are processes described in business processes modeling language (BPML), they can be executed using standard engines. Execution of the processes requires that the input data, information and knowledge are represented using standard formats. It is reasonable to use the same standards for representing the results of measurements processing.
For formal description of data, information and knowledge about initial and processed measurements a hierarchy of information models has been developed [8]. In [6] a set of general classifiers for technologies, methods, algorithms and procedures for measurements processing is proposed. To use the intelligent technologies in the end user applications it is necessary to implement the models and to integrate them into the information models of the applications. For implanting the models it is proposed to use ontological approach as, at first, it has in fact become a standard for describing models of subject domains and, at second, the information models of the applications are commonly described using ontologies.
In the paper a structure of the system of ontologies build according to the models for measurements processing is proposed. Main data mining techniques and models required for measurements processing are enumerated in the second section. In the third section the developed system of ontologies is described. An example of the ontologies adaptation for the subject domain of telemetric information processing (TMI) is given in the fifth section.

2 Models and techniques for measurements processing and analyses

The actual standard of data and information processing and analyses is defined by the JDL model [9]. The JDL model is a general functional model of data and information fusion. The model has five levels: signal level, object level, situation level and level of threats. The highest fifth level is the level of decision making support. Measurements processing and analyses includes three steps: measurements harmonization, integration and fusion. Optionally measurements exploration can be executed at the fourth step. For each of the models levels, the functions and the processes of the levels are defined. The detailed descriptions of the models are given in [10] and the technologies of data harmonization, integration and fusion that provide the implementation of the models can be found in [11]. Input and output parameters of the levels of the functional models are represented using three specialized information models for description of different types of initial measurements and information and knowledge about them: a model of time series of measurements, a model of separate measurements and a combined model of different types of measurements. The description of

each model is given in [3]. Processing of measurements at each level according to the developed technologies assumes application of an a priori defined set of intelligent technologies or separate statistical and data mining methods and algorithms adapted for solving tasks of measurements processing.

The set of intelligent technologies used for measurements harmonization is oriented on processing and analyses of initial binary data streams and the measurements represented in the form of single values or time series that are extracted from the streams. Processing and analyses of initial data streams assumes application of technologies for identification of the structures of the streams and estimation of the quality of the received data. Extracted measurements are transformed into standard formats and described in terms of the dictionary of the subject domain. Harmonization technology uses methods for measurements transformation into different formats, methods based on computing correlation functions, methods based on statistical laws of linguistic distribution, methods for building formalized descriptions of the initial data streams and measurements.

Intelligent technologies oriented on measurements integration include two key technologies: a technology for measurements preprocessing and a technology for preparing measurements for solving applied tasks. The first technology is implemented using algorithms of measurements denoising, removing single and group outliers, filling gaps, removing duplicating values and specialized procedures developed for different types of measurement instruments. The second technology uses methods for estimating compliance of the measurements to requirements of the end user tasks, methods for computing various features of measurements and characteristics of the analyzed objects.

Technologies of data fusion include technologies of extracting information and knowledge from initial measurements, of revealing dependencies in behavior of the measured objects parameters, of grouping measurements, of building grids on the base of separate measurements and of solving separate highly complicated computational tasks. The technology of extracting information and knowledge from measurements is based on algorithms of classification, cluster analyses and segmentation. The technology of revealing dependences applies algorithms of associations mining and building temporal patterns. The technology of measurements grouping is oriented on identifying groups of similar measurements and uses methods of cluster analyses. For the identified groups classes and association rules are defined. The technology of building grids is used to build both regular and non-regular hierarchical grids with various levels of detailing. The list of the computational tasks can include various tasks that are solved at the level of situations or oriented on decision making support.

The list of the technologies and methods given above is aimed to show the multiplicity of the directions of data mining techniques application for processing measurements. The detailed description of each technology one can find in [6]. The data, information and knowledge required to execute the methods and the algorithms directly affect the structure of the information models of measurements and results of their processing and, consequently, the structure of the system of ontologies for measurements processing.

3 A system of ontologies for measurements processing

The proposed interconnected ontologies are aimed to store and to provide data, information and knowledge about measurements and results of their processing. They are developed according to [12] and form the core of the system of ontologies of the subject domain of measurements processing. The system includes 3 main groups of ontologies: ontologies that contain information and knowledge about measurements, ontologies that describe technologies, methods, algorithms and procedures for measurements processing and analyses, and ontologies for representing information and knowledge about objects and situations using measurements of objects parameters. The first group contains the ontologies of time series, of time series segments, of time series features, of time series formal descriptions, of the criteria for the initial measurements and results of their processing estimation. The second group includes ontologies that provide information and knowledge about technologies of measurements processing, applied methods, algorithms and procedures including semantic descriptions of their input and output parameters, conditions of their application, the criteria for estimating results, the history of the methods application as well as other parameters. Ontologies of objects contain information about the structures of objects, their life cycles, functionality, possible interaction, defined regular states and faults. Ontologies of situations define the possible types of situations and provide extended formalized descriptions of situations and the objects involved in the situations.

Different kinds of external ontologies that are required for measurements processing or contain information about related subject domains can be used, for example, ontology of data providers or ontology of statistical distributions. For adaptation to applied subject domains the system can be extended with the specialized ontologies. The set of relations defined for the ontologies is given in Fig. 1.

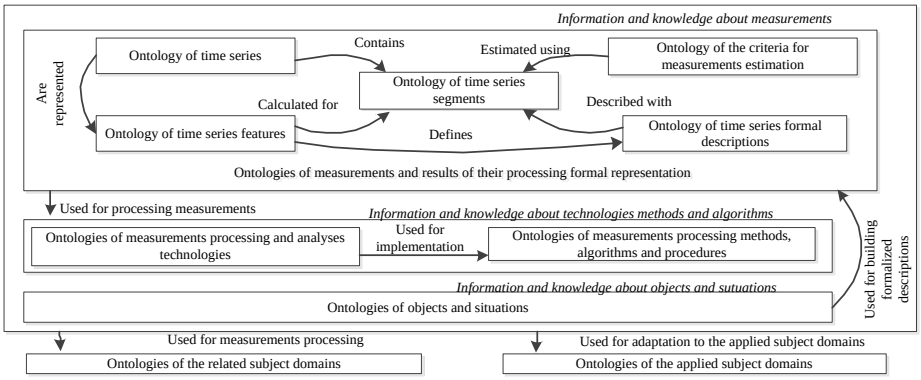


Fig 1. Relations defined for the system of ontologies

A. Description of the ontology of time series. The ontology of time series is aimed to provide information about different types of time series that can be processed. Types are formed according to behavior of time series and consequently define groups of algorithms that one can use for processing time series. The behavior of time series is described using five base features.

Feature 1. According to the types of the objects parameters 3 types of time series of measurements can be defined: functional, signal and constant. Functional time series are represented with continuous functions. For signal time series stepwise behavior is typical. Constant time series do not change in time.
Feature 2. Depending on dynamic of changes of functional time series slow changing time series and fast changing time series can be defined. The first type of time series can be characterized with the frequency spectrum in an interval from 0 up to 20-50 Hz, the second type – up to 2-3kHz or even more.
Feature 3. Depending on behavior, functional time series can be stationary, non-stationary and piece-wise stationary time series. The majority of time series are non-stationary but they contain comparatively long stationary segments.
Feature 4. For slow changing time series existence of gaps in the first and the second derivatives are considered as features.
Feature 5. For functional time series possibility of their description using parametric models is considered. For non-stationary time series a set of parametric models for each of the stationary segments is build. For selecting an appropriate model the models are matched using the least squares method or the method of maximum likelihood estimation.
For defining types of time series for each time series a set of various features is computed and classifiers of the time series types are used. The classifiers can be built on the base of historical data using algorithms for building decision trees [13].
B. Description of the ontology of time series segments. Segments are defined for piece-wise stationary and non-stationary time series. The ontology contains information about possible types of segments that can be observed in a time series. For defining types of segments 2 approaches are proposed. The first approach is based on using an a priori defined set of typical segments that are described in the ontology. To define a type of a segment, similar segments are found in the data base. The data base contains segments that have constant, linear increasing / decreasing, convexly / concavely increasing / decreasing behavior. The data base can be extended with segments that describe specialized behavior of time series typical for the applied subject domain. Specialized segments can be defined by experts or revealed from the historical data. The second approach assumes that for the analyzed segment a set of features is computed. The computed features contain several groups of features that reflect general behavior of the segment, describe the segment without taking into account the local peculiarities of the segment and that are focused on describing all tiny peculiarities of the segment. For defining methods and algorithms for computing features ontologies of methods are used.
C. Description of the ontology of time series features. The ontology is aimed to define features for describing stationary, piece-wise stationary and non-stationary functional time series and segments of time series. The sets of features computed for other types of time series, are fixed. The features can be defined according to the time required for features computing, according to the domain of the time series representation (time, frequency, time-frequency or spatio-temporal domain) and according to information density of the features for the solved task or for the allied subject domain.

The first group of features contains statistical features (median, mode, range, rank, standard deviation, coefficient of the variation, moments including mean, variance, skrewness, kurtosis), measurements frequency, behavior of the curve that corresponds to the time series in the time domain (convexity / concavity of the curve, variability of the curve, the error of the piece-wise constant / piece-wise linear approximation, the error of the approximation using the polynomials of the second and higher degrees, values of the characteristic points, the curvature), entropy, variability of the first derivative. The considered list of features contains commonly list feature, it can be extended or modified. The second group includes feature that consider time series as stochastic processes, in particular, one-dimensional and multi-dimensional distribution functions, one-dimensional and multi-dimensional probability density of the sophisticated processes, the distributions of the probabilities of the sophisticated discrete variables, spectral density. The list of features of the third group that are computed for both initial and transformed time series is given in table 1.

Table 1. Extended set of time series features

Transformation types	Computed features
initial measurements; ranging of values of initial measurements; computation of derivative using the finite difference method; computing of upper and lower envelopes	error of a time series description using a constant / linear / quadratic function for a time series approximation
computation of variation of upper and lower envelopes of a time series	deviation from zero
interpolation using cubic splines	error of interpolation transformation
approximation using a defined function; computation of a curve length	error of approximation transformation using power / exponential / logarithmic / user function
computation of a curve complexity	local complexity, global complexity and weighted complexity
computation of a curve variability	variability indices
computation of the characteristic points of a curve	number of minimums, maximums, intersections with the defined level of the values
computation of a curve curvature	minimum, maximum and median of a curvature
computation of area of a figure that is limited by the curve and the line that connects the edge points [14]	value of an area
computation of the first component using the method of principle components [15]	error of a time series description using a constant / linear / quadratic function for a time series approximation

The alternative approach for building the ontology of the time series features is proposed in [16]. It is based on computing linear, non-linear and other features. For defining linear features measures based on the computing of linear correlation, frequency parameters of the time series and autoregressive models are used. To nonlinear features refer 19 features. Definition of measures for these features assumes computation of nonlinear correlation and of time series dimension and complexity, building nonlinear models of time series.

D. Ontology of time series formal descriptions. The ontology is used for building formal descriptions of stationary, piece-wise stationary and non-stationary functional

time series. Descriptions are built according to the computed features of the time series. The time series can be described using adaptive and non-adaptive approaches [17]. Adaptive approach assumes computing coefficients of piece-wise constant and piece-wise linear approximations, coefficients of singular decomposition and building symbolic representations of time series. In order to build non-adaptive descriptions one can use such features as coefficients of wavelet transformations, of time series spectral representation, results of piece-wise aggregate approximation. Depending of time series complexity one or several descriptions can be built.

E. Description of the ontology of criteria for initial measurements and results of their processing estimation. In the ontology 3 groups of criteria for initial measurements are considered. The first group allows one to estimate measurements using knowledge about the object / environmental area which parameters are measured, the second group − using results of matching new data with historical data, the third group − using specialized procedures selected according to the types of the processed measurements and applied methods. The criteria of the first group are usually defined by experts and / or producers of the measurement instruments. They are represented as a set of features for which admissible intervals for measured values are given. The second group of criteria is based on computing distances between the analyzed measurements or their features and measurements that were acquired earlier in similar conditions. The third group of the criteria includes criteria that estimate separate measurements and sets of measurements, separate time series and their groups. The criteria significantly depend on the solved tasks. The examples of the criteria are uniqueness, accuracy, consistency, completeness, timeliness, actuality, interpretability, relatedness to other data.

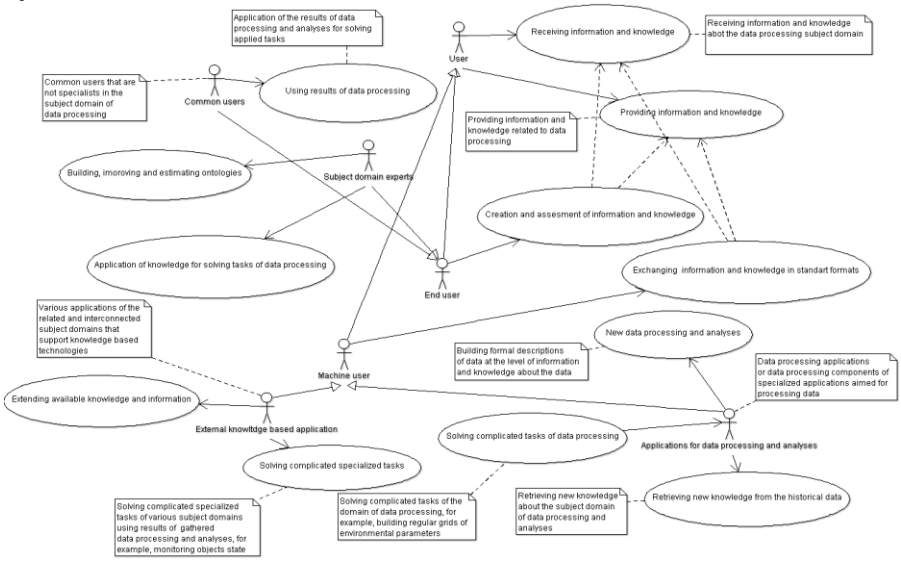


Fig 2. Use case diagram for the system of the ontologies for measurements processing

Results of measurements processing are estimated twice: just after measurements are processed and at consequent stages of their processing and analyses. Both stages assumes application of the procedures of revealing contradictions of the acquired results

with available information, of comparing results received using different methods, of comparing results with results of historical data processing, of comparing results of separate measurements and separate time series processing with the results of joint analyses, of computing complex features on the base of separate features. An example of criteria for cluster analyses methods can be found in [18].
The described above system of ontologies but can be used for solving tasks in intelligent applications specialized for measurements processing by experts and common users and by different external applications. The use case diagram for the proposed system of ontologies is given in Fig. 2.

4 Application of the system of ontologies for TMI processing

The developed set of ontologies for measurements processing was adapted for processing TMI [19] received from remote space objects. A hierarchy of the solved tasks is given in Fig. 3.

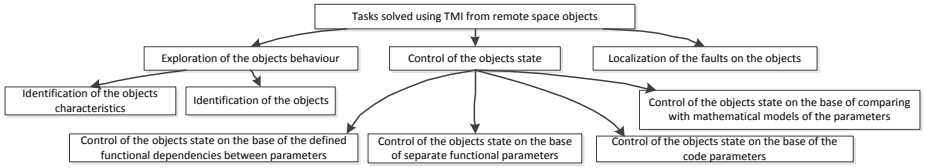


Fig 3. Ontology of the tasks

Table 2. Time series of measurements of specialized parameters

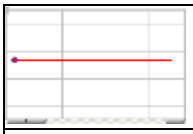
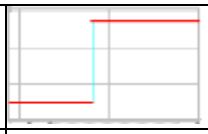
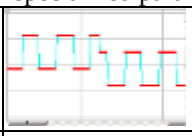
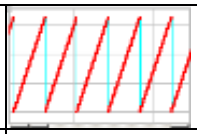
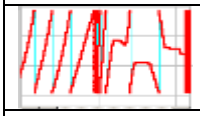
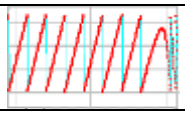
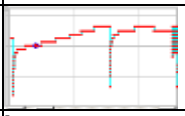
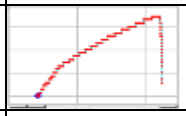
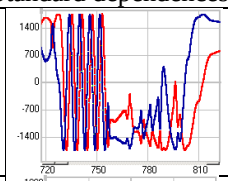
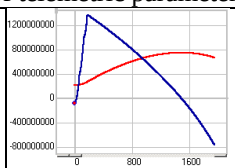
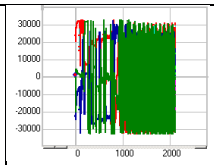
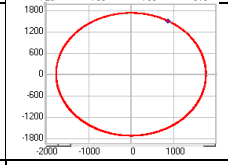
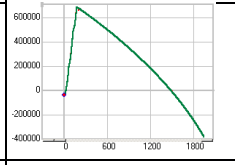
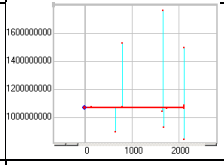
			
constant	code	meander	counter
			
mantissa	order	lower part	upper part

Table 3. Standard dependences of telemetric parameters

Example of the initial data			
Dependency graphical representation			
Dependency	pairs of sine and cosine $x^2 + y^2 = 1$	integro-differential pairs $x' - y = 0$	elements of the matrix of the coordinates transformation

Adaptation required extension of the ontology of the types of times series, the ontology for representing dependences in objects parameters and the ontology of methods and algorithms for measurements processing. A set of types of time series was extended with the types aimed to describe measurements of specialized parameters (table 2). The set of features for the specialized types are defined in [20]. The standard dependencies of telemetric parameters include pairs of sine and cosine, the integro-differential pairs and elements of the matrix of the coordinates transformation (table 3). The upper level ontology of methods and algorithms for TMI processing is given in Fig.4. Several branches of the ontology are detailed in Fig. 5-7.

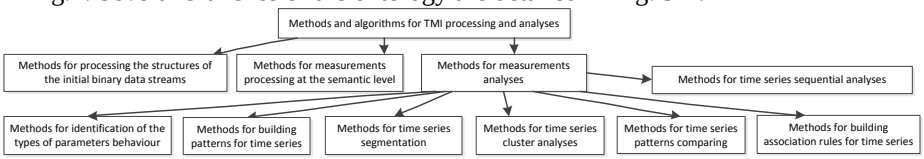


Fig 4. Ontology of methods and algorithms for TMI processing and analyses

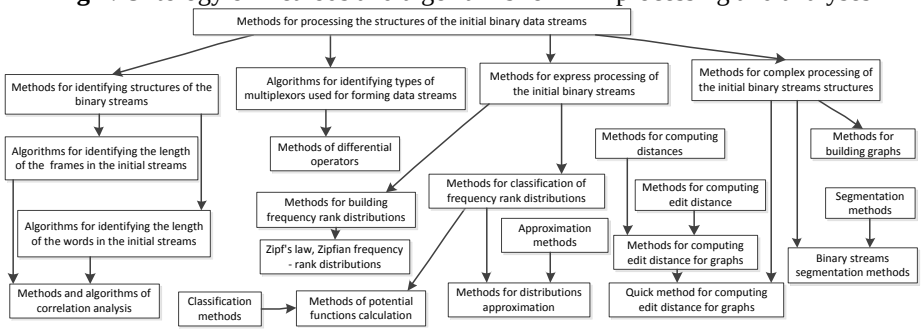


Fig 5. A fragment of the ontology of methods for processing structures of binary streams

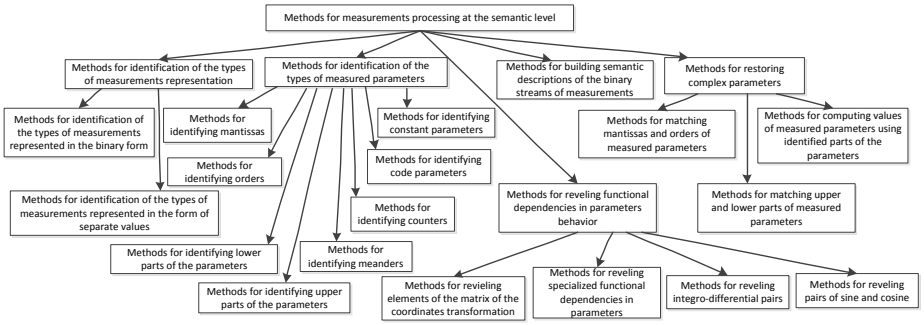


Fig 6. A fragment of the ontology of methods for measurements processing at the semantic level

The system of the ontologies was implemented in a number of the applications oriented on processing TMI from space objects in the delayed mode that are successfully used for about ten years already. The description of the developed systems and the examples of their application can be found in [6, 21].

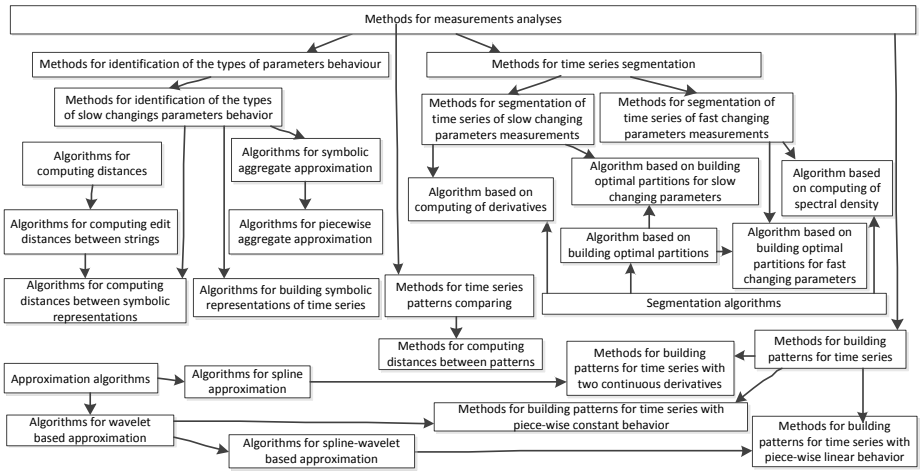


Fig 7. A fragment of the ontology of methods for measurements analyses

5 Case Study

The control of the space objects state using code parameters assumes analyses of the time points at which the values of the parameters changed. These points correspond to the moments of execution of commands on the controlled objects. In table 4 a subset of code parameters for three different objects of one type are given. For each parameter the time points of their values change are defined.

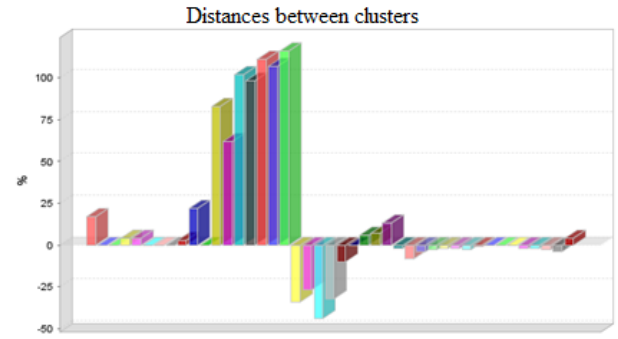
Table 4. The time of the values change points of the code parameters

№	*PRMp*	*KND*	*SC*	*Ki*	*PRMb*	*OPKi*	*OHKi*	*ST*	*KZ*
1	362789	344936	348956	350428	350429	359539	359535	361435	361746
2	453563	464518	468542	470111	470113	479124	479121	481018	481328
3	190444	201398	205418	206915	206917	216025	216018	217898	218208
	KP4b	*KP4c*	*KP4d*	*KP4e*	*KK4a*	*KK4b*	*KK4c*	*KK4d*	*KK4e*
1	361479	361478	361483	361483	361944	361943	361940	361944	361955
2	481061	481061	481062	481063	481475	481476	481476	481477	481477
3	217941	217941	217942	217943	218355	218356	218357	218357	218358
	KD1a	*KD1b*	*KD1c*	*KD1d*	*KD1e*	*PRK*	*KD3b*	*KD3c*	*KD3d*
1	362327	362373	362366	362387	362388	362789	363040	363040	363042
2	481930	482010	481990	481991	481970	482372	482623	482605	482606
3	218832	218833	218870	218891	218871	219252	219482	219497	219482
	KD3e	*GK*	*KD5a*	*RPbc*	*RPcd*	*RPde*	*RPeb*	*VOGb*	*VOGc*
1	363042	363042	363295	363300	363299	363307	363300	363330	363329
2	482645	482653	482906	482910	482908	482915	482909	482933	482927
3	219494	219504	219746	219748	219746	219755	219750	219777	219775
	VOGd	*VOGe*	*VNNb*	*VNNc*	*VNNd*	*VNNe*	*KP*	-	-
1	363330	363330	363332	363331	363332	363330	363356	-	-
2	482919	482930	482940	482938	482938	482939	482950	-	-
3	219778	219777	219781	219780	219784	219783	219791	-	-

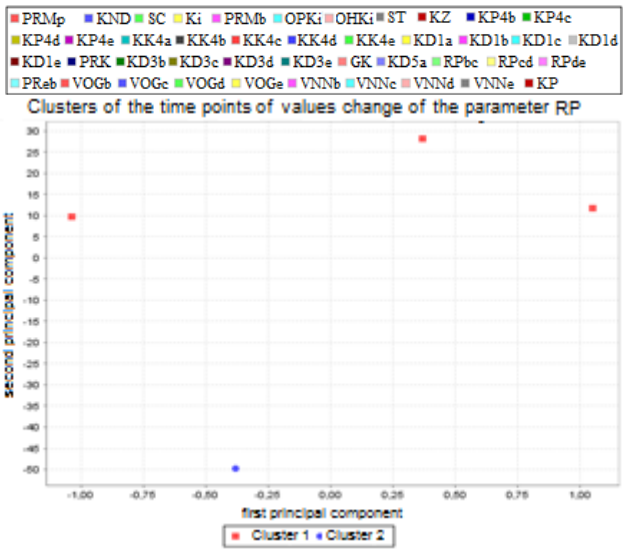
The time points of the values change were processed using data mining techniques, in particular, statistical and cluster analyses methods. The results of building clusters

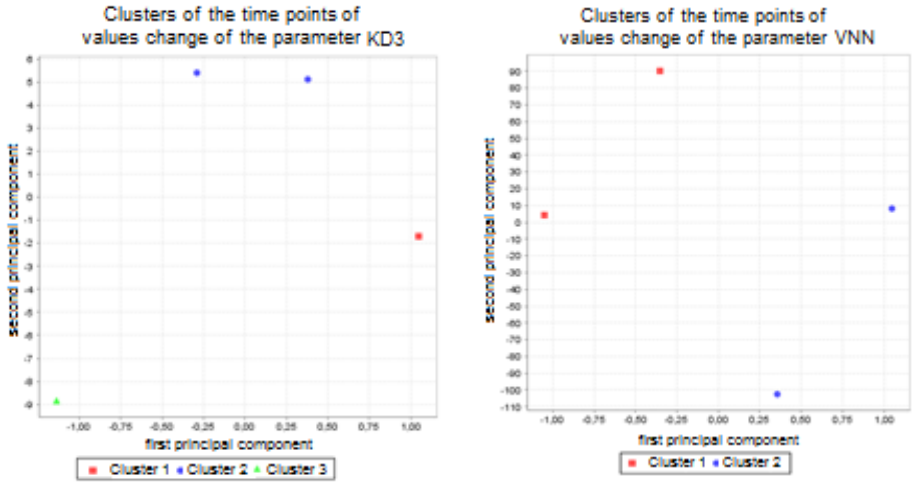
of objects using all parameters showed that the behavior of the first object differs significantly from the behavior of the second and the third objects. The first object is the only element of the first cluster. The second and the third objects form the second cluster. The differences between the clusters are represented in the form of a histogram (Fig. 8 a). The order of the parameters in the histogram is the same as in the table 4. The cluster analyses of similar parameters of different blocks of the objects that have equal construction (the name of the block to which the parameters refer is written in small letters after the name of the parameter) revealed deviations from the normal behavior for the parameters RPde (the time points of the disconnection of the spherical locks between blocks 'b' and 'e' differ from the time points defined for the same parameter between other blocks), KD3 (the time points of the contacts breaking of blocks 'b' and 'd' differ from the time points defined for the parameter for blocks 'c' and 'e'), VNN (the time points of the output of the tooth for blocks 'd' and 'e' differ from the time points defined for blocks 'b' and 'e') (Fig. 8 b-d). The clusters in Fig. 8 are represented in the feature space build using the principal component method [22].

a)



b)





c) d)

Fig 8. Application of Data Mining techniques for processing time points of the values change of code parameters

6 Conclusion

In the paper a system of ontologies required for processing and analyzes of various objects parameters measurements represented in the form of time series or single values is presented. The structure of the ontologies and the relations between the ontologies that link them into a system are defined. For each of the ontologies a detailed description is provided and the relations with external ontologies are enumerated.

The proposed system of the ontologies has the following distinguishing features:

- the system allows one to solve the tasks of measurements processing taking into account the peculiarities of the processed data and the solved tasks;
- multiple technological solutions for measurements processing based on application of intelligent methods and algorithms can be implemented using the considered set of ontologies;
- the structure of the system of the ontologies and of the separate ontologies is simple and can be easily extended and modified if new methods are developed or new types of measurements are defined;
- information and knowledge represented in the form of ontologies can be interpreted both by experts and machines and can be multiply used;
- the system of ontologies can be easily adapted to different subject domains if ontological descriptions of the domains are available.

Further development of the described system of ontologies assumes detailing the ontologies on the base of knowledge, acquired as a result of operating of the developed applications for telemetric information processing. A set of applications for other subject domains is going to be developed and approved.

References

1. Steinberg A.N. Foundations of Situation and Threat Assessment, Handbook of Multisensor Data Fusion, D. Hall, M. Liggins, J. Llinas (eds.), LLC Books (2008).
2. Steinberg, A.N. ; Rogova, G. Situation and context in data fusion and natural language understanding. Proceedings of 11th FUSION, Cologne (2008).
3. Vitol A., Zhukova N., Pankin A. Adaptive multidimensional measurements processing using IGIS technologies. Proceedings of the 6th International Workshop on Information Fusion and Geographic Information Systems: Environmental and Urban Challenges, St. Petersburg (2013)
4. Pankin A., Vodyaho A., Zhukova N. Operative Measurements Analyses in Situation Early Recognition Tasks. Proceedings of the 11th International Conference on Pattern Recognition and Image Analyses, Samara (2013)
5. Zhukova N. Method for adaptive multidimentional meas-urements processing based on IGIS technologies. Proceedings of the 11th International Conference on Pattern Recognition and Image Analyses, .Samara (2013)
6. Vitol A., Deripaska A., Zhukova N., Sokolov I. Technology of adaptive measurements processing. SPbSTU «LETI», Saint-Petersburg (2012)
7. Browne P. JBoss Drools Business Rules. Packt Publishing (2009)
8. Vitol A., Zhukova N., Pankin A. Model for knowledge representation of multidimensional measurements processing results in the environment of intelligent GIS. Proceedings of the 20th International Conference on Conceptual Structures for Knowledge Representation for STEM Research and Education, Mumbai (2013)
9. Steinberg A., Bowman C., White F. Revisions to the JDL Data Fusion Model. Sensor Fusion: Architectures, Algorithms, and Applications. Proceedings of the SPIE, vol. 3719 (1999)
10. Zhukova N. Harmonization, integration and fusion of multidimensional measurements of technical and natural objects parameters in monitoring systems [in Russian]. Izvestiya SPbETU "LETI", vol 2, Saint-Petersburg (2013)
11. Popovich V., Voronin M. Data Harmonization, Integration and Fusion: three sources and three major components of Geoinformation Technologies. Proceedings of IF&GIS, St. Petersburg (2005)
12. http://www.w3.org/
13. Quinlan R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
14. Feng S., Kogan I., Krim H. Classification of curves in 2D and 3D via affine integral signatures. Acta Applicandae Mathematicae, vol 109, issue 3, Springer, Nitherlands (2010)
15. Chang K., Ghosh J. Principal curve classifier - a nonlinear approach to pattern classification. Proceedings of Neural Networks, Anchorage (1998)
16. Kugiumtzis D., Tsimpiris A. Measures of Analysis of Time Series (MATS): A MATLAB Toolkit for Computation of Multiple Measures on Time Series Data Bases. Journal of Statistical Software, vol. 33, issue 5 (2010)
17. Lin J, Keogh E., Lonardi S., Chiu B. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego (2003)
18. Halkidi M., Batistakis Y., Vazirgiannis M. Clustering Validity Checking Methods. ACM Sigmod Record 31(2,3) (2001)
19. Nazarov A., Kozyrev G., Shitov I. et al.: Modern Telemetry in Theory and in Practice. Training Course [in Russian]. Nauka i Tekhnika, St. Petersburg (2007)

20. Vasiljev A., Vitol A, Zhukova N. Detecting the symantic structure of the group telemetric signal [in Russian]. SPbSTU «LETI», Saint-Petersburg (2010)
21. Vasiljev A., Geppener V.,Zhukova N.,Tristanov A.,Ecalo A. Automatic control system of complex dynamic objects state on the base of telemetering information analysis [in Russian]. 8th International Conference on Pattern Recognition and Image Analysis: New Information Technologies, vol.2, No.4 (2007)
22. Jolliffe I. Principal Component Analysis. Springer, 2nd ed. (2002)

Система онтологий для приложений обработки данных на основе техник анализа данных

Александр Водяхо[1], Наталья Жукова[2]

[1] Санкт-Петербургский государственный электротехнический университет, Санкт-Петербург, Россия
[2] Санкт-Петербургский институт информатики и автоматизации Российской академии наук, Санкт-Петербург, Россия
{aivodyaho,nazhukova}@mail.ru

Аннотация. В статье описана система онтологий, спроектированных для приложений, ориентированных на решение проблем распознавания и оценки ситуаций на основе результатов обработки и анализа данных. Основное внимание сосредоточено на проблемах обработки измерений от различных объектов с параметрами, представленными в виде временных рядов. Рассмотренные приложения обрабатывают данные при помощи знаний, извлечённых из исторических данных при помощи техник анализа данных. Такие приложения очень зависят от базы знаний, представляющей собой систему онтологий. Представленная система онтологий является множеством онтологий верхнего уровня, для которых разработаны способы решения задач в одной или нескольких предметных областях.

Ключевые слова: представление знаний, анализ данных, слияние данных, обработка измерений, распознавание и оценка ситуаций.

Logic-Mathematical Apparatus of Data Processing Used in Information Technology of Web-Portal Development

Svitlana Bevz

Vinnytsia National Technical University, Vinnytsia, Ukraine
svbevz@rambler.ru

Abstract. The paper suggests the improved logic mathematical apparatus, used for development of computer systems, the given apparatus enables to unify the description of information models and determine the existing links between databases tables. Models, serving as the base for information technology of automated web-portal design, have been developed, using the given apparatus. The creation of the structure and algorithm of information technology for automated data processing has been performed.

Keywords: information models; logic-mathematical apparatus; information technology; web-portal.

1. Introduction

Today automated data processing is one of the most important tasks in information technology (IT) industry. There exists quite a good number of IT for solution of the problems, dealing with the data processing in information systems (IS) [1-3]. However, data management in IS with geographically-distributed structure is quite time-consuming task and requires a specialized approach to provide the integration of information space of complex hierarchical computer system subjects.

For monitoring, managing, processing and analysis of IS data various models, methods and modeling tools are used [4-8]. Wide functionality and classification of separately taken methodologies, however, does not allow to solve the complex of problems of data consolidation on the Web portal, in particular, formation of hierarchical structures, created using specific methods, for example containing recursion for monitoring of catalogs and elements or groups and subgroups of users.

To unify and extend the functionality of the existing methods in [9], the author of this article offered a logical mathematical apparatus for data processing, which enables to automate the process of information systems design by means of models transformation in the user interface of Web portal. The task of the development of IT for processing of data of Web Portal hierarchy structure requires extending of functionality area and application of prior developed logic-mathematical apparatus.

Taking into account current trends of web-based information technologies development [10, 11], for testing and promotion of research results of Masters, Postgraduates,

applicants for the scientific degrees and young scientists the Institute of Graduate, Postgraduate and Doctoral Studies (InGPDS) of Vinnytsia National Technical University (VNTU) developed a young scientists web-portal as an interactive environment, oriented to the filling of online resources of virtual scientific-educational space.

Solution of urgent problem of information technology development, intended for processing of young scientists portal data, requires the usage of logic-mathematical models for automation and unification of data management methodology.

The aim of research is to increase operation efficiency of information system used for monitoring, data analysis and processing. The object of the research is information technology of automated processing data at young scientists portal. The subject of research are logic-mathematical models of web portal data processing.

To achieve this aim the following problems should be solved:

– improvement of the logic -mathematical apparatus of data processing;

– construction of information model of web portal data set ;

– development of logic-mathematical models for analysis and processing of the portal elements data;

- development of the information technology architecture for web portal data processing;

- development and introduction of information technology software in higher establishments.

2. Logic-mathematical apparatus of automated data processing

In the process of web portal operation its information content is constantly updated and refreshed by the user, new problems arise, they often require non-trivial approaches to their solution, for instance, change of data structure in the process of system operating, change of user interface, change of design patterns. In order to consolidate the information and data processing and for the solution of the above-mentioned problems, the usage and further development of logic-mathematical apparatus developed in [9] is suggested.

Table 1 presents twelve information models for the automation of data processing, with the description of their components.

Table 1. Information models of data processing, designation of operations and symbols

№ пп	Model	Designation	Description
1	$D = A\ [A_1,\ \ldots,\ A_n]$	$A_k \in A,\ k = \overline{1,\ n}$ – attributes of table A	Projection of table A
2	$D = A(h(A))$	$h(A)$ – logic condition using attributes of table A	Sample of table A
3	$D = A\langle q(A,\ B)\rangle B$	$q(A,\ B)$ – logic condition of tables A and B consolidation	Rigid union of tables A and B
4	$D = A\langle\langle q(A, B)\rangle B$	$\langle\langle q(A, B)\rangle$ – logic condition of tables A and B external consolidation	External unification of tables A and B (each record of table B is united with the record set of relation A).
5	$D = A\begin{bmatrix} A_1,\ \ldots,\ A_n, \\ f(G_1), \ldots,\ f(G_m) \end{bmatrix}$	$f(G_i)$ – function of aggregate attributes usage	The use of agregate functions (avg, sum, $count$, max, min – functions of mean, total, quantity, maximum and minimum value).
6	$D\lVert P_1, \ldots, P_{n_P}\rVert = A\,(h(A, P))$	$P = \lVert P_1, \ldots, P_{n_P}\rVert$ –parameters of the model; $h(A, P)$ – logic condition for A and P	Model of parameters use
7	$D = A\ [A_1, \ldots, A_n]$ $(w(A, \{B'\}))$; $B' = B\,[\,f(B_m)\,](h(B, A))$	B' – model of nested query; $w(A, \{B'\})$ – condition with a nested query	Model of nested query use in the condition
8	$D = A\ [A_1, \ldots, A_n, \{B'\}]$	B' – model of nested query	Model of nested query use in the attributes
9	$D\lVert P_1, \ldots, P_{n_P}\rVert = A\,(h(A, P))$ $\to Z\{B'\} \to X$; $B' = B(h(B, P, Z))$; $R = Z \vee X$	B' – projection of table; $Z\{B'\}$ – internal model; X – result of internal model; $h(B, P, Z)$ – condition of the retrieve from the set B and parameters P	Model of parameters transfer from internal model

10	$D\left\|P_1,\ldots,P_{n_P}\right\| =$ $\{B\left\|P'_1,\ldots,P'_{n_B}\right\| \to$ $\left[B_1,\ldots,B_{n_{B'}}\right]\}\left[B_{n_{B1}},\ldots,B_{n_{B2}}\right]$	$P=\left\|P_1,\ldots,P_{n_P}\right\|$ – set of parameters; $B=\left[B_1,\ldots,B_{n_B}\right]$ – the result of the nested model	Model of parameters use in nested query
11	$D=A\left[H_1(A_1),\ldots,H_m(A_m),\right.$ $\left.f(A_{m+1}),\ldots,f(A_n)\right];$ $H_k(A_k)=$ $=\begin{cases} S_0(A_k)=A_k,\ A_k\notin SA; \\ S_m(A_k)\neq H_j(A_k); \\ A_k\in SA; \\ j\neq k;\ m=\overline{1,n_S-1}; \end{cases}$ $A_k\neq G_m,\ A_k\in A,\ k=\overline{1,\ m};$ $G_r\in A,\ SA\subset A,\ r=\overline{m+1,\ n}$	$S_j(A_i),\ j=\overline{1,n_s}$ – function of ordinal sorting of attributes; $S_0(A_i)=A_i$ – function of attributes usage without sorting	Model of data sorting and grouping
12	$D=rec\left(C\left(h(C)\right)\langle g(C,B)\rangle\right.$ $\left.\{B\to\left\|B_1,...,B_k\right\|\}[sum(B_k)]\right)$	$h(C)$ – initial recursion condition, $g(C,B)$ – condition of subsequent element of recursion , B – the internal query of recursion ; $[sum(B_k)]$ –summation of recursive queries results by attribute B_k	Model of recursion

Model of parameters usage in the nested request and recursion model, used in computer system for automatic processing of hierarchical data structures, traditionally found IS of web-portals are added to ten models, developed before [9].

Therefore, logic-mathematical apparatus has been improved and supplemented by two new models. The construction of information models of automated data processing of the computer systems in particular – young scientists web-portal is realized using the suggested logical-mathematical apparatus.

3. Informational model of data set

The basis of young scientists web-portal, as in any automated web-based system is a database. Software orientation of web portal, which is integrated in the unified automated information system of document management and monitoring of educational process of Masters' training [3], provides information and analytical possibilities of modern web-based system with distributed structure and uses the database of the existing system.

Data bank structure of young scientists' portal in rather simplified form is shown in Fig.1. It contains four main units: users, directories and elements, models and relationships, characteristics of the interface. Module of models and relationships plays a key role in managing portal objects and subjects.

The diagram contains designations of tables and their attributes. They will be used for construction of logic-mathematical models for portal data processing.

4. Logic-mathematical models of information technology

Let us consider practical application of logical-mathematical apparatus, models of which are presented in Table 1. We will construct information models of data management and processing for young scientists portal.

Portal authorization is performed during user identification by means of login *Xlo* and password *Xps* using model of projection and retrieval:

$$Xid = U[Uid]\ ((Ulo = Xlo) \wedge (Ups = Xps)). \quad (1)$$

In case of a successful identification (Xid > 0) information system (IS) determines the user code Xid. Guest login (Xid = 0) restricts the rights of portal users.

Users belonging to the administrators group (*Gid =4*) is represented by the model using agregate function to calculate the number of tuples:

$$Xadm = G\ <Gid=Agr>A\ [count(Gid)]\ ((Gid =4) \wedge (Aus=Xid)), \quad (2)$$

which determines the parameter of the system *Xadm*, computing the amount of set *A* records of target groups and portal user code portal. Similarly user identification of belonging to other groups of the portal members is performed.

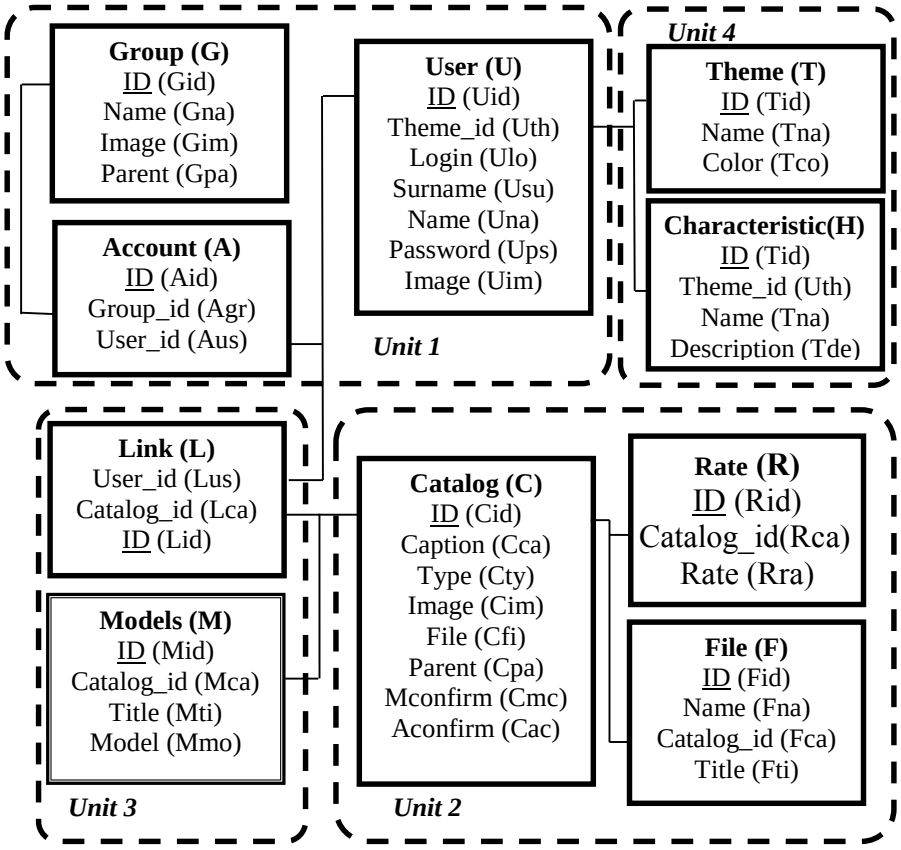


Fig. 1. Databank of young scientists' portal

Since each element and the catalogs, descriptions of which are stored in the set *C,* may refer simultaneously to several users of the portal, then we write the model of all elements and catalogs of the portal *E*, which is formed by the union of the sets of catalogs and elements *C* and interlinks with users *L* by means of the model of rigid combination:

$$E = C < (Cid=Lca) > L. \quad (3)$$

Provisional table of catalogs and their elements is formed on request of guest viewing of *Vg*, confirmed by data administrator using sample model combination:

$$Vg=C\ (Cac =1). \quad (4)$$

Revision of user's information portal *Vu* displays personal elements and catalogs, published by him on a Web page, as well as data objects, confirmed by administrator, as the union of data sample models:

$$Vu = E((Eus = Xid) \wedge (Eac = 0)) \vee C(Cac = 1). \quad (5)$$

Administrators display all the elements and catalogs *Va*, which are confirmed by the participant of the portal – author of information objects, according to the sample model:

$$Va = C(Cmc = 1). \quad (6)$$

Information regarding catalogs and elements of the portal is expedient to store in one relation table of database. Tuples of their data differ by attribute *Cty*, which equals 1 for the catalogs, and for elements – *2*. Number of subsidiary elements *Vk* for display of k^{th} catalog:

$$Vk = rec(((Cty = 1) \wedge (Cid = k)) < Cid = Rpa >$$

$$\{C[Cpa, count(withid)] \rightarrow |Rpa, Rcnt||\}[sum(Rcnt)]). \quad (7)$$

Recursion is used in (7), which complements the logic-mathematical apparatus (see Model 11 of table 1), using the parameters *Rpa, Rcnt* – code of the parent element and the amount of its subsidiary elements, respectively.

Review of m^{th} element of catalog *m* displays file *Vm* and a list of hyperlinks *Vn* in accordance with the following data sample models:

$$Vm = C[Cfi](C_{id} = m); \quad Vn = F[Fna, Fti](Fca = m). \quad (8)$$

Some catalogs of the portal are characterized by rating of users scientific achievements. Total rating takes into account all the achievements of s^{th} participant of Vs portal, presented in his catalogs and elements and is determined by the model of data retrieval and aggregate:

$$Vs = R < (Eca = Rca) \wedge (Eus = s) > E[sum(Rra)]. \quad (9)$$

Formulas (1)-(9), developed on the common methodological base are information models of logical-mathematical apparatus for web-portal data processing in accordance with the above-mentioned structure (see Fig. 1).

We will carry out the construction of information technology of information processing, relying on the above-mentioned models.

5. Information technology of data processing

Based on the suggested logical-mathematical models, methods of processing data that realize the function of automated design of SQL requests of information portal are developed, this enables to change rapidly the structure of the data and parameters of the computer system in the process of its operation and provides the necessary degree of integrity of information database, and also improves the efficiency of data processing of the portal due to reduction of time for data consolidation and formation of users' queries.

IT of young scientists' portal construction is based on the developed models and methods of data processing.

Let us consider the architecture of information technology and basic principles of its functioning. Fig. 2 shows the architecture of the information technology for automated processing of portal data.

Management of information system interface is carried out design patterns of user interfaces that on the base of style sheet and description of Web pages structural blocks realize combination of the data model processing results and user interface elements, which, in its turn, allows to process efficiently data sets using standardized software functions.

Information technology is implemented in accordance with modular principles of libraries shown in Fig. 2, which provide wide functional possibilities of the portal. Module of models formation provides information support of the process of the development of logical-mathematical models for functional support of the system, processing and data analysis and creating account forms. To save objects, interface parameters, and interlinks with subjects database server MySQL is used.

Created information technology architecture, unlike the existing technologies allows:

– use one and the same element both for publication on a personal website that is automatically generated for the portal user and in the tree of its catalogs accessible to Internet community, or to the community of researchers, scientific schools;

– perceive one and the same user as a participant of various groups at a fixed moment of time and at various moments of time, taking into retrospective of catalogs and their content;

– use one and the same element of the portal, taking into consideration time retrospective for several authors, enabling to save the resources of the disc space and time of data processing;

–perform data analysis and computation of scientific rankings both for separately taken portal user and for groups of users, for example, scientific school department, institute.

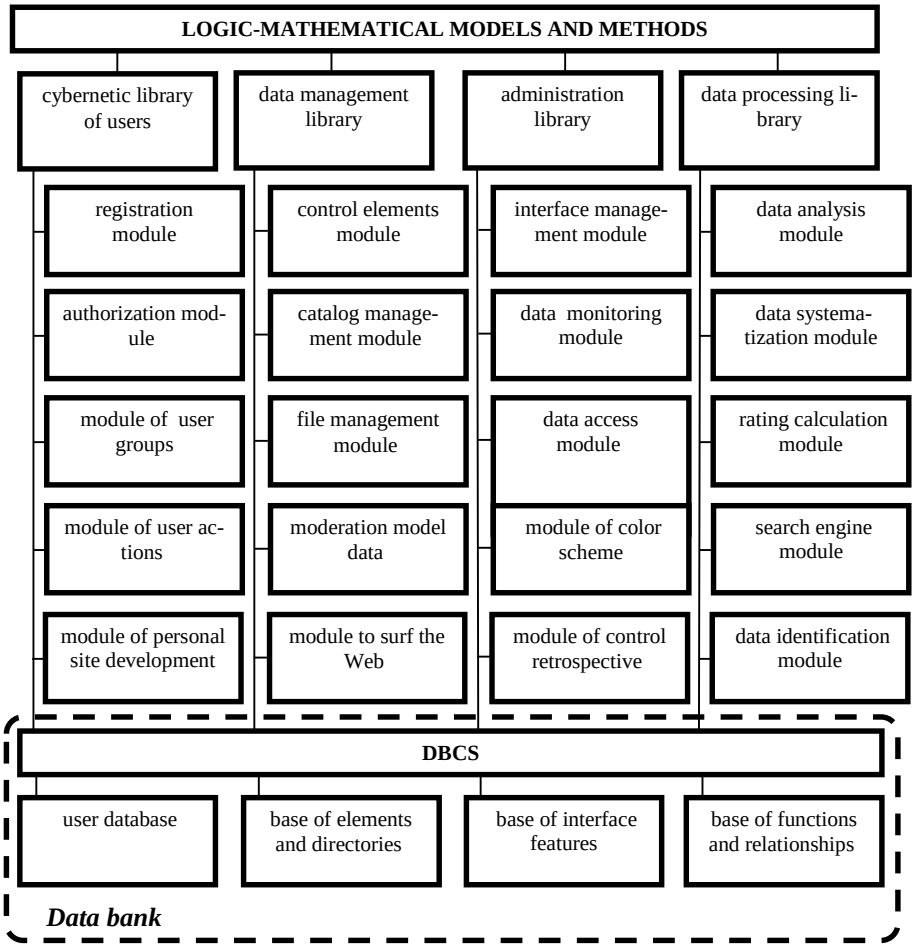


Fig. 2. Information technology architecture

Using the suggested methods of automated processing and data analysis algorithm of information technology that performes the formation of information space for young scientists portal and provides its functionality has been developed: analysis and data processing, elements and catalogs control, user groups management, administration, creation of personal users websites by data publishing and editing directly in the system of formation and browsing the web document, loading of text and graphic information, web-surfing of multimedia files, posted on public servers.

On the base of algorithm of portal functioning software modules for information technology realization are developed. These modules perform monitoring and analysis

of information and enable to improve the effectiveness of automated data processing in information system.

6. Practical implementation of data processing information technology

Efficiency and performance of the developed models, algorithms and methods of information technology for data processing is proved by their program implementation and application at Vinnytsia National Technical University. Young scientists portal for (http://inmad.vntu.edu.ua/portal/) in 2014 started its work and today registration of new users, catalog creation and filling them with information resources is performed. The portal interface is illustrated in Fig. 4.

The code of the portal is written in PHP language using MySQL database server and Apache web server. Design and test version of the portal were developed in 2013 in the environment Visual Studio Ultimate 2012 using the database server SQL Server 2012 Enterprise Edition [12].

Further use of the suggested information technology for automated data processing and accumulation of information resources for web-portal will allow to perform calculation of the scientific schools and individual scientists trajectory of development to determine the optimal vector for accumulation of scientific knowledge with the criterion of optimality in various scientific practical fields.

Today the scientific community of Ukraine is open for communication with scientists and researchers in other countries and realizes joint projects, development and implementation of innovative technologies, in particular, in the field of education. Information, regarding new research achievements is of great importance for training of highly qualified scientific stuff. Presentation of the research results on the portal pages increases the motivation of young scientists, enhancing the efficiency of their scientific research.

7. Conclusions

Logic-mathematical apparatus of models of automated data processing representation was further developed in the given research. By means of this apparatus it becomes possible in a simple and understandable form describe the parameters of data analysis, processing, aggregation, and take into account existing relationships between tables of relational database. Models of parameters usage in the nested request and recursion are added to logic-mathematical apparatus, thereby expanding the scope of this apparatus and automate management and data processing functions of hierarchical structures. By

means of the suggested logic-mathematical apparatus the construction of application fields information models – portal is realized on the common methodological base.

By means of a computer system based on information models of data processing automated generation of necessary SQL requests and construction of information tables using template design of the page is carried out, that enables to improve the efficiency of web-portal data processing.

New technology of information processing is developed. It differs from existing systems by logic-mathematical models of data formation and allows to consolidate and arrange portal elements, taking into account the data structure of the given information system and allows to enhance the efficiency of resource management.

Practical implementation of developed information technology of data processing automation in the program resource – young scientists portal, has been illustrated, on the example of the portal put into operation at Vinnytsia National Technical University.

In future we plan to continue the research of functional possibilities of logical-mathematical apparatus in the sphere of data analysis and management of their processing, and also extend the sphere of logic-mathematical models application on other areas for information technology implementation.

References

1. Edelhauser Eduard. Management information systems. A case study over the last eight years in the Romanian organizations [Electronic resource] / Eduard EDELHAUSER, Lucian LUPU–DIMA // Database Systems Journalvol. III, no. 3/2012.
2. Grady Booch The Unified Modeling Language User Guide: Second Edition// Grady Booch, James Rumbaugh, Ivan Jacobson. – Addison Wesley Professional. – 2005. – 496 p. – ISBN: 0-321-26797-4.
3. Mokin B. I. Information and communication technology of automated monitoring and educational process management of master training /B.I. Mokin, V.B. Mokin, S.V. Bevz, S.M. Burbelo // Information technologies and learning tools. – Vol.23. – №3. – 2011.
4. Sandro Etalle Logic Programming // Sandro Etalle, Miroslav Truszczunsky / 22nd International Conference ICLP 2006, Seattle USA, August 2006 Proceedings. – ISBN: 3-540-36635-0.
5. Lex De Haan Applied Mathematics for Database Professionals / Lex De Haan, Toon Koppelaars // New York : Apress. – 2007. – 376 p.
6. Khomonenko, A. D., Tsygankov, V. M., Maltsev, M. G. Databases: textbook for higher educational establishments: edited prof. A. D. Khomonenko, Sankt-Petersburg: KORONA-print, 2004, 736 p., 4-th Edition.
7. O. L. Berezko, A. M. Peleschyshyn and P. I. Zhezhnych, "Conception of Modern University Website Development: Case Study of Lviv Polytechnic National University," in Proceedings of the 5th International Conference of Young Scientists "Computer Science and Engineering": CSE-2011, November 24-26, 2011, Lviv, Ukraine. Lviv: Vydavnytstvo Lvivskoi politekhniky Publ., 2011. pp. 370-373.

8. Maier D. Logic and Lattices for Distributed Programming / D.Maier, N. Conway, W. Marczak, P. Alvaro and J. Hellerstein. Proceedings of the ACM Symposium on Cloud Computing (SoCC '12). San Jose, October 2012.
9. Romanyuk A. N. Building Automation of the Computer Systems of Management Reporting / A. N. Romanyuk, S. V. Bevz, S. M. Burbelo // International IEEE SIBCON. – 2011. – Tomsk. – 978-1-4577-1070-4/11/2011 IEEE. – P. 136-139.
10. Banday M. Tariq Web Portal for Kashmir Tourism Industry: Design Guidelines // Sprouts / Proceedings of 4th J & K Science Congress 12th to 14th Nov, 2008, University of Kashmir. – ISSN: 1535-6078.
11. Sadeh Tamar, Walker Jenny Library portals: toward the semantic Web // The Emerald Research. – New Library World. – Volume 104. – No 1184/1185. – 2003. – pp. 11-19. – ISSN 0307-4803.
12. Mokin V. Web-portal of young scientists of VNTU / V. Mokin, S. Bevz, V. Voytko, S. Burbelo, and others // Proceedings of I International Internet-Conference "Problems and technologies in training scientific personnel with higher qualification under conditions of innovative development of society". – Vinnytsia.: VNTU, 2013. – p. 83-86. – ISBN: 978-966-641-551-9.

Логико-математический аппарат обработки данных для использования в информационных технологиях при разработке Web-порталов

Свитлана Бевз

Винницкий национальный технический университет, Винница, Украина
svbevz@rambler.ru

Аннотация. В статье предложен развитый логико-математический аппарат, используемый при разработке информационных систем. Предложенный аппарат позволяет унифицировать описание информационных моделей и определить существующие связи между таблицами базы данных. Модели в основе информационной технологии, используемой при автоматическом построении Web-портала, построены при помощи предложенного аппарата. Создана структура и алгоритм технологии для автоматической обработки данных.

Ключевые слова: информационные модели; логико-математический аппарат; информационные технологии; Web-портал.

Semantic Methods of Structuring Mathematical Content and Open Scientific E-Journals Management Systems[1]

Alexander Elizarov, Evgeny Lipachev, Denis Zuev

Kazan (Volga Region) Federal University
{amelizarov,elipachev,dzuev11}@gmail.com

Abstract. The paper discusses the approach to automate the processing of electronic mathematical documents and their transformation into semantic documents. Structuring electronic storage of periodical issues in mathematics and multi-volume works conferences was performed.

Keywords: information and communication technologies, information resources, technologies of the Semantic Web, electronic scientific collections, metadata, mathematical notation, electronic publications, Open Journal System.

[1] This study was supported by the Russian Foundation for Basic Research (project No 12-07-97018) and Russian Scientific Foundation for Humanities (project No 14-03-12004).

References

1. Rocha, E.M., Rodrigues, J.F.: Disseminating and preserving mathematical knowledge. In: Borwein, J.M., Rocha, E.M., Rodrigues, J.F. (eds.). Communicating Mathematics in the Digital Era. pp. 3–21. A K Peters, Ltd. (2008)
2. Miner, R.: The importance of MathML to mathematics communication. Notices of the AMS, 52, 532–538 (2005)
3. Elizarov, A.M., Lipachev, E.K., Malakhaltsev, M.A.: Web technologies for mathematician: MathML bases. Practical guidance. Fizmatlit, Moscow (2010)
4. Elizarov, A.M., Lipachev, E.K., Hohlov, Yu. E.: Semantic methods of structuring mathematical content providing enhanced search functionality. Information Society. No 1-2, pp. 83-92 (2013)
5. Kohlhase, M.: Using $L^{A}T_{E}X$ as a semantic markup format. Math. Comput. Sci. 2, pp. 279304. Birkh¨auser Verlag Basel/Switzerland (2008)
6. Lange, C., Kohlhase, M.: A mathematical approach to ontology authoring and documentation. In: Carette, J. et al. (eds.). Calculemus/MKM 2009, LNAI 5625, pp. 389–404. Springer-Verlag, Berlin, Heidelberg (2009)
7. Nevzorova, O., Zhiltsov, N., Zaikin, D., Zhibrik, O., Kirillovich, A., Nevzorov, V., Birialtsev, E.: Bringing math to LOD: a semantic publishing platform prototype for scientific collections in mathematics. Lecture Notes in Computer Science. 8218, pp. 379–394. Springer (2013)`
8. Friedl, J.: Mastering regular expressions. O'Reilly Media Inc. (2008)
9. Elizarov, A., Zuev, D., Lipachev, E.: Open scientific e-journals management systems and digital libraries technology. http://ceur-ws.org/Vol-1108/paper13.pdf

Семантические методы структурирования математического контента и открытые системы управления электронными научными журналами

Александр Елизаров, Денис Зуев, Евгений Липачев

Казанский (Приволжский) федеральный университет
{amelizarov,elipachev,dzuev11}@gmail.com

Аннотация. Обсуждается подход к автоматизации процесса обработки научных электронных документов и их преобразования в структурированные документы. Акцент сделан на особенностях обработки математических текстов. С помощью сервисов, созданных по предложенной методике, выполнено структурирование достаточно большого по объему электронного хранилища, содержащего выпуски периодического журнала по математике и многотомных трудов конференций.

Ключевые слова: информационно-коммуникационные технологии, информационные ресурсы, технологии Семантического Веба, электронные научные коллекции, метаданные, математическая нотация, электронные публикации, Open Journal System.

Методика построения функции принадлежности для классификации изображений на основе гистограмм яркости

Иван Посохов, Ирина Сергеевна Логунова

Магнитогорский государственный технический университет им. Г.И. Носова»,
Магнитогорск, Россия
posohof@gmail.com, logunova66@mail.ru

Аннотация. Приведены особенности изображений образцов используемых при оценке качества полуфабрикатов и готовой продукции в металлургической промышленности. Выдвинута гипотеза о возможности разделения изображений на три класса. Построена методика классификации изображений по гистограмме яркости. Методика была опробована в ходе вычислительного эксперимента.

Ключевые слова: изображение, гистограмма, принятие решения

1 Краткий теоретический анализ проблемы исследования изображений металлургической продукции

Современное промышленное производство выдвигает новые требования к системам управления многостадийными производствами. Эти требования обусловлены внедрением новых приоритетных направлений, определенных государственной политикой в России. Одно из таких направлений – развитие информационно-телекоммуникационных технологий, которые являются неотъемлемой частью автоматизированных систем управления (АСУ) производством крупных промышленных предприятий. Использование новых модулей АСУ для многостадийных производственных процессов способствует повышению эффективности функционирования агрегатов и обеспечивает снижение доли продукции пониженного качества [1, 2].

При разработке и внедрении новых модулей, дополняющих существующие АСУ производства, появляется необходимость использования графической информации, получаемой в ходе оценки качества готовой продукции и полуфабрикатов.

В области теории и практики использования графической информации и принятия решений в условиях АСУ производств накоплен значительный положительный опыт. Вопросы получения, обработки и сегментации изображений отра-

жены в трудах зарубежных и российских исследователей. Труды [3-5] определили развитие математической теории в области обработки графической информации.

Однако, несмотря на проведенные исследования и значительное число публикаций в области обработки графической информации, остаются актуальной проблема: отсутствие комплексных методик, позволяющих выполнять автоматическую обработку изображений, характерных при формировании базы данных информации и качестве металлургической продукции.

В сложивших условиях возникает необходимость в разработке автоматической гибкой системы обработки изображений, включающей отделение фона и объекта исследования, а также идентификации в пределах выделенного объекта исследования элементов, соответствующих нарушению сплошности образца.

2 Определение цели и задач исследования

Учитывая проблемы оценки исходных изображений для металлургической продукции, была определена цель исследования как совершенствование существующих методов и средств анализа обработки графической информации о качестве металлургической продукции для ее последующего использования в управлении системой многостадийного производства непрерывнолитой заготовки.

Для достижения цели авторами работы решаются задачи:

- проведение теоретико-информационного и теоретико-множественного анализа многостадийного производства непрерывнолитой заготовки и определение возможности и необходимости внедрения в систему управления производством эргатического модуля анализа изображения серного отпечатка;
- построение комплекса методик, включающего алгоритмы анализа графической информации на основе классификации изображений по гистограмме яркости, автоматического отделения фона от объекта исследования, классификацию и идентификацию объектов на изображении, соответствующих нарушению сплошности заготовки;
- построение методики экспертной оценки информации и принятие решений с использованием результатов идентификации объектов на изображении.

В рамках представляемой статьи представлена методика анализа графической информации на основе классификации изображений по гистограмме яркости.
Научную новизну, в отличии от известных методов классификации изображений, составляет методика, использующая характеристики бимодальной гистограммы, такие как положение минимального и максимального значения яркости.

3 Характеристика набора исходных изображений для оценки качества

По результатам пассивного экспериментального исследования, включающего сбор графической информации о качестве непрерывнолитой заготовки была сформирована база данных, содержащая 32 серных отпечатка. Изображения представляют собой оцифрованные серные отпечатки непрерывно литой заготовки (рис. 1).

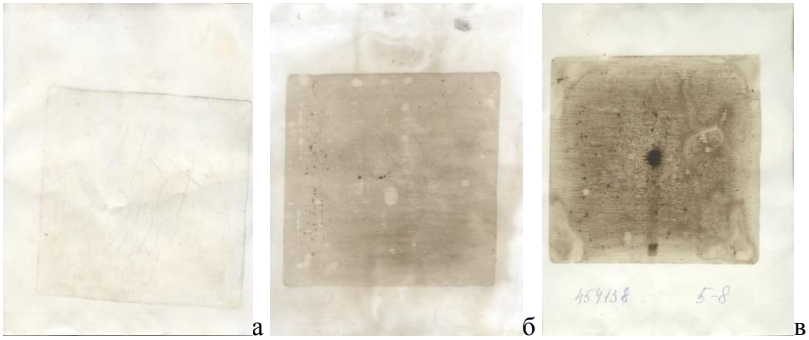


Рис. 1. Изображения серных отпечатков поперечных темплетов непрерывнолитой заготовок квадратного сечения: а – типовое изображение группы А с незначительным отличием яркости фона и объекта исследования; б - типовое изображение группы В со средним отличием яркости фона и объекта исследования; в - типовое изображение группы С с резким отличием яркости фона и объекта исследования

Оцифровка серного отпечатка производится путем его сканирования. Так как размер серного отпечатка в среднем составляет 175×230 мм, то для его оцифровки применятся сканер формата А4 с разрешающей способностью не менее 300 точек на дюйм. Результирующие изображения в среднем имеют разрешение 4200×5500 точек.

В ходе визуального анализа всех изображений было выявлено две особенности:

- изображение отпечатка всегда содержит объект исследование (непосредственно изображение поперечного сечения заготовки размером 100×100 мм) и окружающий фон, причем положение поперечного сечения (изучаемого объекта) является неопределенным (рис. 2);
- все изображения можно разделить на три группы по отношению яркости объекта исследования и фона изображения. Типовые изображения каждой групп в порядке возрастания этого отношения приведены на рис. 1.

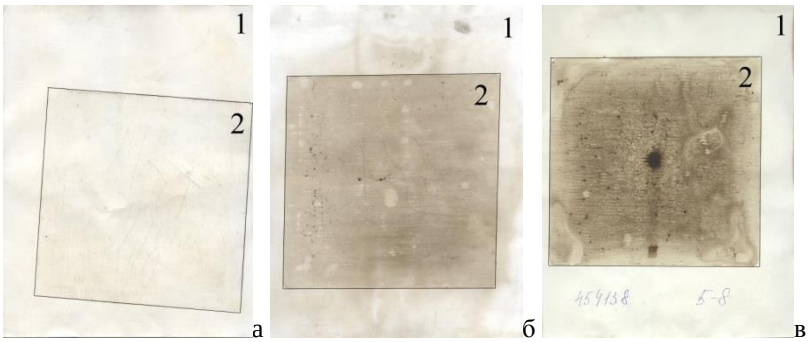


Рис. 2. Структура исследуемых изображений в терминах «Объект исследования – Фон»: а,б,в – номером 1 обозначен фон, номером 2 – объект исследования

Для сокращения объемов интерактивного определения области непосредственного исследования принято решение о построение алгоритма автоматического выделения этой области. Для автоматического поиска объекта на изображении многими авторами [9-10] предлагается использовать методы, основанные на сегментации изображений. Наиболее простым представителем методов сегментации является пороговая обработка. Данный метод применим для изображений содержащих известное число однородных по яркости классов точек, распределение вероятностей которых унимодальны. Кроме этого, граничные участки между замкнутыми областями должны занимать сравнительно небольшую площадь изображения.

Предлагается также применить методы сегментации на основе выделения границ (контуров). Методы требуют, чтобы между областями изображения существовал резкий перепад яркости, что характерно не для всех исходных изображений.

Поэтому авторы работы предлагают классифицировать изображения по гистограмме яркости преобразованного полутонового изображения.

4 Гистограммы яркости и преобразование изображения в полутоновое

Авторами работы предварительно был проведен анализ гистограмм яркости для каждого канала пространства RGB[1] и гистограмм яркости преобразованного полутонового изображения. Исходные растровые изображения серных отпечатков являются полноцветными и представлены в цветовой модели RGB с глубиной цвета 24 бита на пиксель и приводятся к полутоновым изображениям с глубиной цвета 8 бит на пиксель. Под полутоновыми изображениями в контексте работы

[1] В RGB модели каждый цвет представляется красным, зеленым и синим первичными основными цветами (компонентами).

понимаются изображения в оттенках серого. Каждый пиксель такого изображения содержит информацию об интенсивности (яркости). Суммарное число возможных градаций яркости для 8-битового полутонового изображения составляет $2^8 = 256$. Минимальное значение яркости – 0 соответствует черному цвету, максимальное – 255 соответствует белому цвету.

Основная стратегия преобразования полноцветного изображения в полутоновое заключается в использовании принципов фотометрии для сопоставления яркости изображения в оттенках серого и яркости исходного цветного изображения [6]. Яркость результирующего пикселя вычисляется как взвешенная сумма трех значений интенсивности модели RGB. Веса цветовых компонент выбраны в соответствии со стандартом ITU-R BT.709, который учитывает особенности человеческого восприятия, большая чувствительность к зеленому цвету и меньшая к синему:

$$Y = 0{,}2126 \cdot R + 0{,}7152 \cdot G + 0{,}0722 \cdot B,$$

где Y – яркость результирующего пикселя; R, G, B – значения цветовых компонент пикселя исходного изображения [7].

Гистограммой цифрового изображения называется дискретная функция

$$h(r_k) = n_k,$$

где r_k – k-ый уровень яркости; n_k – количество пикселей изображения с яркостью r_k [4]. Для 8-битового изображения k изменяется в пределах [0; 255].

На рис. 3 – 5 приведены гистограммы для типовых изображений серных отпечатков, приведенных на рис. 1 каждого класса. На рис. 3 – 5 введено обозначение: N – количество пикселей на изображении заданного канала, тыс шт.

Для каждой группы изображений были получены гистограммы, имеющие характерные отличия в форме:

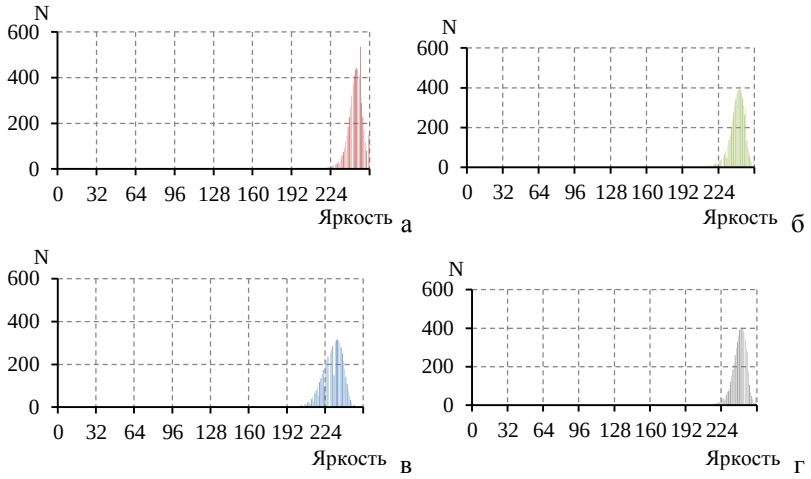


Рис. 3. Гистограммы для изображения группы А: а – гистограмма для красного канала; б – гистограмма для зеленого канала; в – гистограмма для синего канала; г – гистограмма полутонового изображения

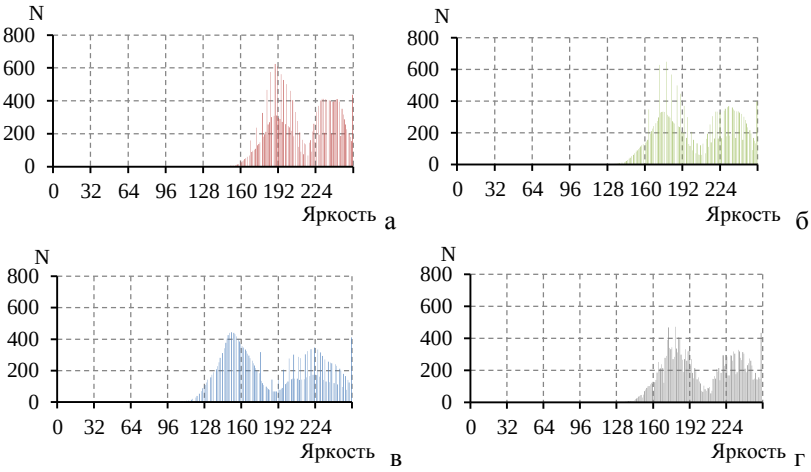


Рис. 4. Гистограммы для изображения группы В: а – гистограмма для красного канала; б – гистограмма для зеленого канала; в – гистограмма для синего канала; г – гистограмма полутонового изображения

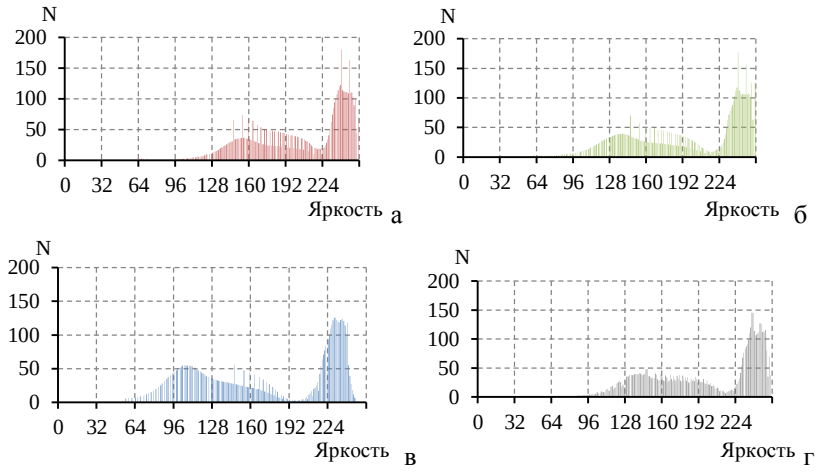


Рис. 5. Гисторамммы для изображения группы С: а – гисторамма для красного канала; б – гистограмма для зеленого канала; в – гистограмма для синего канала; г – гисторамма полутонового изображения

- гистограммы изображений группы А унимодальны, так как фон и объект исследования имеют сравнительно одинаковую яркость. Значимые уровни яркости занимают узкую (10÷15% в диапазоне [215, 255]) полосу, показывая, что изображения имеют низкий контраст. Полоса смещена к левому краю, что свидетельствует о высокой яркости изображений;
- для гистограмм изображений группы В характерно наличие двух выраженных пиков, сопоставимых по площади, так как фон и объект исследования имеют различную яркость. Значимые уровни яркости занимают полосу шириной 40÷50% в диапазоне [140, 255] – изображение имеет нормальную контрастность;
- гистограммы изображений группы С также имеют два пика, однако левый пик занимает более широкую полосу и меньшее максимальное значение. Фон и объект исследования имеют существенно отличающуюся яркость. Ширина полосы значимых уровней яркости 70% в диапазоне [80, 255], что соответствует высококонтрастному изображению.

Таким образом, оценка полученных гистограмм по каждой группе показала, что:

- гистограмма полутонового изображения является достаточной для классификации изображений, так как отклонение положений порога и максимумов канальных гистограмм по сравнению с гистограммой полутонового изображения составляет не более 15%;
- начальное предположение о разделении выборки изображений на три группы подтверждена результатами вычислительного эксперимента и появилась необходимость в разработке методики идентификации формы гистограммы изображении серного отпечатка.

5 Методика идентификации формы гистограммы яркости серного отпечатка

Для отнесения гистограммы к одному из классов изучены методы нахождения порогового значения по источникам [3-5,8,11]. Согласно этим источникам пороговое значение – это величина яркости, относительно которой гистограмма делится на две части.

Разделение гистограммы на две части позволяет найти максимы яркости каждой части. Поиск порогового значения яркости основан на использовании метода Оцу [8]. Метод позволяет разнести пиксели изображения на два класса, рассчитывая такой порог, чтобы внутриклассовая дисперсия была минимальной. Такая дисперсия выражается через взвешенную сумму дисперсий двух классов:

$$\sigma_{\omega}^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t) \quad (1)$$

где веса ω_i – это вероятности двух классов, разделенных порогом t; σ_i^2 – дисперсия этих классов.

Вероятность для каждого уровня интенсивности рассчитывается как:

$$p(t) = \frac{n(t)}{N} \quad (2)$$

где *n(t)* – количество пикселей изображения с яркостью *t*; N – общее количество пикселей изображения.

Автор метода [8] доказал, что минимизация дисперсии внутри класса равносильна максимизации дисперсии между классами

$$\sigma_b^2(t) = \sigma^2 - \sigma_{\omega}^2(t) = \omega_1(t)\omega_2(t)[\mu_1(t) - \mu_2(t)]^2 \quad (3)$$

где μ_i – среднее арифметическое класса.

После применения метода Оцу и нахождения значения порога яркости методика предполагает нахождение положения максимумов дискретной функции для каждой части гистограммы. Блок-схема методики классификации изображений низкой контрастности на основе гистограммы яркости приведена на рис. 6. Поиск выполнен простым перебором значений дискретной функции гистограммы в каждой ее части.

В блок-схеме приведенной на рис. 6 подпроцесс «Получение изображения» включает в себя чтение полноцветных изображений серных отпечатков непрерывнолитой заготовки с носителя и размещение его в памяти в виде массива пикселей. Подпроцесс «Построение канальных гистограмм яркости» выполняет построение гистограммы яркости для каждого из трех каналов в цветовом пространстве RGB, красного, зеленого и синего. В подпроцессе «Изображение к полутоновому» происходит преобразование изображения из полноцветного в полутоно-

вое, используя алгоритм BT709. За ним следует подпроцессы «Построение гистограммы полутонового изображения» для построения гистограммы изображения полученного на предыдущем шаге. Действия, входящие в подпроцессы «Поиск порогового значения» и «Поиск положений максимумов», представлены на блок-схеме (рис. 7).

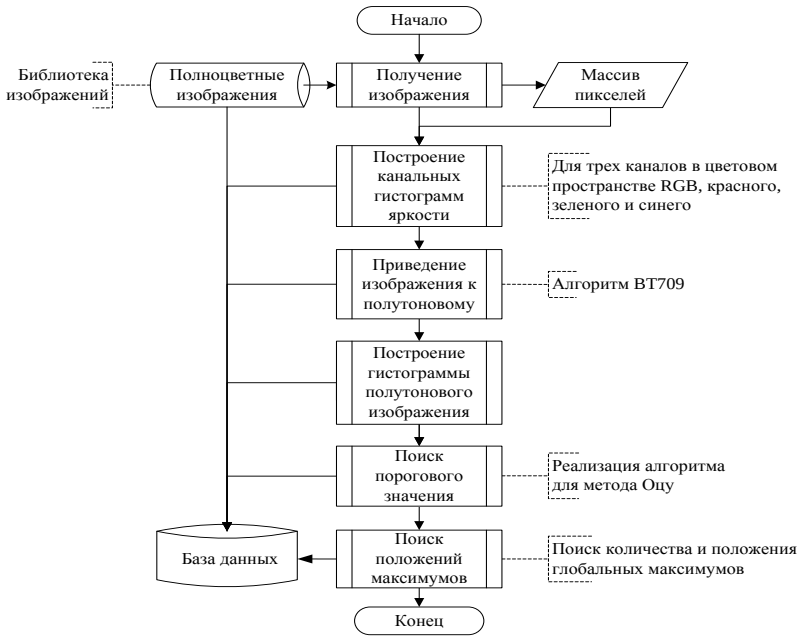


Рис. 6. Блок-схема методики классификации изображений на основе гистограммы яркости

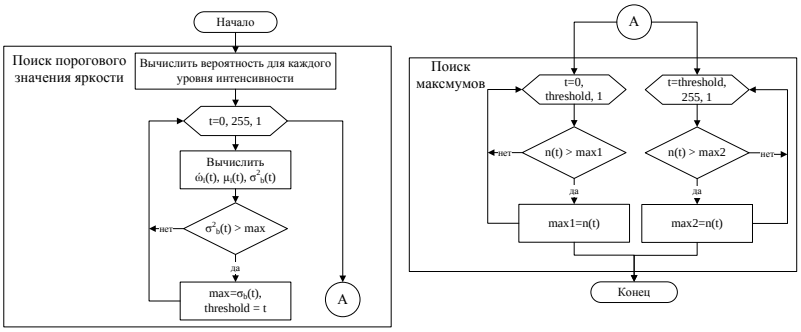


Рис. 7. Блок схема поиска порогового значения яркоркости и максимумов дискретной функции гисторраммы изображений

В результате применения методики формируется база данных, которая содержит следующие сведения: исходное изображение, полутоновое изображение,

дискретные функции для построения гистограмм для красного, зеленого и синего канала, а также полутонового изображения, пороговое значение и значения максимумов в обеих частях гистограммы.

6 Результаты тестирования методики классификации изображений на основе гистограммы яркости

Для разработанной методики были проведены вычислительные эксперименты для всех собранных изображений. Для изображений, приведенных на рис. 1 были получены результаты, приведенные в табл. 1.

Таблица 1. Результаты оценки порога яркости и максимальных значений дискретной функции для тестовых изображений

Изображение	Красный канал			Зеленый канал		
	T_r	M_{1r}	M_{2r}	T_g	M_{1g}	M_{2g}
Рис. 1а	219	212	248	224	201	242
Рис. 1б	214	189	255	178	205	255
Рис. 1в	199	154	240	194	146	240
Изображение	Синий канал			Полутоновое		
	T_b	M_{1b}	T_b	M_{1b}	T_b	M_{1b}
Рис. 1а	227	226	227	226	227	226
Рис. 1б	190	151	190	151	190	151
Рис. 1в	177	146	177	146	177	146

В табл. 1 введены обозначения: T_r, T_b, T_g, T_{gs} – значение дискретной функции для порога яркости для каждого канала и полутонового изображения; M_{1r}, M_{1b}, M_{1g}, M_{1gs} – значение максимума слева от порога яркости дискретной функции гистограммы для каждого канала и полутонового изображения; M_{2r}, M_{2b}, M_{2g}, M_{2gs} – значение максимума справа от порога яркости дискретной функции гистограммы для каждого канала и полутонового изображения; r, b, g – индексы для обозначения цветовых каналов соответственно красного, синего и зеленого; gs – индекс соответствия полутоновому изображению.

Таблица 2. Результаты классификации полутоновых изображений серных отпечатков поперечного сечения непрерывнолитой заготовки

Группа	Количество изображений	Значение показателей					
		Порог яркости		Максимум слева		Максимум справа	
		мин.	макс.	мин.	макс.	мин.	макс.
A	4	216	223	193	211	239	254
B	21	193	231	147	218	233	254
C	7	217	240	145	239	236	247

В табл. 2 приведены результаты классификации изображений, включенных в общую базу данных на основе разработанной методики.

Таким образом, для принятия решения об отнесении изображения к одной из групп (A, B или C) можно построить функцию принадлежности адаптивного вида:

$$R = \sum_{i=0}^{3} i \cdot \left(\left(T_{i\,\min} \le T \le T_{i\,\max} \right) \wedge \left(m_{i\,\min} \le m \le m_{i\,\max} \right) \wedge \left(M_{i\,\min} \le M \le M_{i\,\max} \right) \right), \quad (4)$$

где R – множество возможных решений, состоящее из четырех элементов {0, 1, 2, 3}, значение которого соответствуют группам изображений: 1 – группа A; 2 – группа B; 3 – группа C; 0 – группа, объединяющая изображения не входящие ни в одну из групп классификации; T, m, M – количественные характеристики гистограммы изображения, выбранного для классификации: порог, максимум слева и максимум справа соответственно; T_{imin}, T_{imax}, m_{imin}, m_{imax}, M_{imin}, M_{imax} – эмпирические границы диапазона порога, максимума слева и максимума справа, определенные на основе эмпирического исследования и адаптируемые при обучении системы принятия решения о классификации изображений.

7 Заключение

Таким образом, авторами по результатам визуального анализа была выдвинута гипотеза о возможности разделения изображений серных отпечатков непрерывнолитой заготовки на три класса. Для классификации изображений была построена методика классификации изображений по гистограмме яркости, учитывающая ее основные характеристики – значения порога яркости, максимум слева и справа.

Построенная методика была опробована в ходе вычислительного эксперимента, по результатам которого выполнено построение адаптивной функции принадлежности изображении к выбранным классам. Адаптация функции возможна в реальном времени при расширении базы данных изображений серных отпечатков в автоматическом режиме. В результате проведенного вычислительно эксперимента показано, что нет оснований для отклонения гипотезы о разделении исходных изображений на три класса.

Список литературы

1. Logunova, O.S. Integrated system structure of intelligent management support of multistage metallurgical processes / O.S. Logunova, I.I. Matsko, I.A. Posochov. – Vestnik of Nosov Magnitogorsk state technical university, 2013. – № 5. – Pp. 50 – 55.
2. Логунова, О.С. Система интеллектуальной поддержки процессов управления производством непрерывнолитой заготовки: монография / О.С. Логунова, И.И. Мацко, И.А. Посохов. – Магнитогорск: Изд-во Магнитогорск. гос. техн. ун-та им. Г.И. Носова, 2013. – 175 с.

3. Шапиро, Л. Компьютерное зрение / Л. Шапиро, Дж. Стокман. – М.: БИ-НОМ. Лаборатория знаний, 2006. – 752 с.
4. Гонсалес, Р. Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – М.: Техносфера, 2005. – 1072 с.
5. Прэтт, У. Цифровая обработка изображений: пер. с англ. / У. Прэтт. – М. : Мир, 1982. – Кн. 1. – 312 с.
6. Poynton, C. Rehabilitation of gamma / Charles Poynton. – Photonics West'98 Electronic Imaging. International Society for Optics – 1998.
7. A Standard Default Color Space for the Internet – sRGB / M. Stokes, M. Anderson, S. Chandrasekar, R. Motta., 1996. URL: http://www.w3.org/Graphics/Color/sRGB.html (дата обращения 06.01.2014)
8. Otsu, N. A Threshold Selection Method from Gray-Level Histograms / Nobuyuki Otsu. – IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, 1979, pp. 62-66.
9. Методы автоматического обнаружения и сопровождения объектов. Обработка изображений и управление / Б.А. Алпатов, П.В. Бабаян, О.Е. Балашов, А.И. Степашкин. – М.: Радиотехника, 2008. – 176 с.
10. Волков, В.Ю. Выделение прямолинейных кромок на зашумленных изображениях / В.Ю. Волков, Л.С. Турнецкий, А.В. Онешко. – Информационно-управляющие системы. – 2011. – № 4. – С. 13–17.
11. Ridler, T.W. Picture thresholding using an iterative selection method / T.W Ridler, S. Calvard. – IEEE Trans. System, Man and Cybernetics – 1978, SMC-8: 630-632.

The Method of Constructing the Membership Function to Classify Images Based on Histograms

Ivan Posokhov, Oksana S. Logunova

Nosov Magnitogorsk State Technical University, Magnitogorsk, Russia
posohof@gmail.com, logunova66@mail.ru

Abstract. Enumerated the features of image samples used in assessing the quality of semi-finished and finished products in the metallurgical industry. Hypothesized the possibility of dividing the image into three classes. Introduced the image classification method based on the histogram. Method has been tested in the computational experiment.

Keywords: image, histogram, decision-making.

Применение модели векторной авторегрессии для анализа потребления электроэнергии

Нургуль Маматова

НИУ ВШЭ, Москва, Россия
nur_gul90@mail.ru

Аннотация Данная статья посвящена электроэнергетической отрасли, анализируется объем потребления электроэнергии населением и промышленными предприятиями России. Экономические факторы, которые влияют на производство электроэнергии исследованы методом векторного авторегрессионного анализа VAR, взаимосвязь между переменными определена с помощью теста Грэнджер, также рассмотрены сценарии реакции потребления электроэнергии на экзогенные шоки, связанные с изменением исследуемых факторов.

Ключевые слова: эконометрический анализ, тестирование на причинность по Грэнджеру, анализ импульсных откликов.

1 Введение

Исследование потребления электроэнергии очень важно, так как от правильного снабжения необходимым количеством электроэнергии зависит обеспечение необходимым количеством электроэнергии населения, а также уровень производства, что в свою очередь является важным для развития экономики в целом. К примеру, по причине активной политики снижения потребления электроэнергии в России, к 2014 году планируется ввести социальные нормы потребления электроэнергии. Подобное наблюдается уже у многих стран в сфере энергопотребления, где призвано снизить к 2020 году энергоемкость производства на 20-40% [1]. Поэтому важно рассмотрение реакции потребителей в случае увеличения цен на электроэнергию, так как речь идет об уровне благосостояния населения, а также необходимо определение степени влияния изменения тарифов и других факторов на объем потребления электроэнергии.

Потребление электроэнергии зависит от различных факторов, по эмпирическим исследованиям экономистов, выделены такие важные факторы, влияющие на потребление электроэнергии (Y) как производство электроэнергии (X), численность населения (P), инвестиции в основной капитал (I), ВВП (G) и тарифы на электроэнергию (T).

2 Результаты исследования потребления электроэнергии

Векторная авторегрессия (VAR) – это модель динамики нескольких временных рядов, в которой текущие значения этих рядов зависят от прошлых значений этих же рядов. Модель предложена К. Симсом как альтернатива системам одновременных уравнений, которые предполагают существенные теоретические ограничения.

Изначально был проведен предварительный анализ переменных на стационарность с применением расширенного теста Дикки-Фуллер (ADF-test). Для стационарных переменных сделан тест на причинность по Грэнджеру и определена взаимосвязь между переменными, которые исследовались попарно, а также были выделены эндогенные и экзогенные переменные для проведения анализа VAR.

По результату исследования выделены две эндогенные переменные тарифы на электроэнергию (T) и производство электроэнергии (X). Важно отметить, что по результатам регрессионного анализа, эти переменные вошли в модель как значимые переменные. Лаговые значения VAR модели были определены основываясь на критериях LR, FPE, AIC, SC и HQ, и свидетельствуют об оптимальности модели с одним лагом. Временные ряды трехмерной VAR модели для переменных имеют следующий вид:

$$
\begin{aligned}
Y_t &= \alpha_{10} + \sum_{i=1}^{\rho} \alpha_{11} Y_{t-i} + \sum_{i=1}^{\rho} \alpha_{12} X_{t-i} + \sum_{i=1}^{\rho} \alpha_{13} T_{t-i} + U_{1t} \\
X_t &= \alpha_{20} + \sum_{i=1}^{\rho} \alpha_{21} Y_{t-i} + \sum_{i=1}^{\rho} \alpha_{22} X_{t-i} + \sum_{i=1}^{\rho} \alpha_{23} T_{t-i} + U_{2t} \qquad (1) \\
T_t &= \alpha_{30} + \sum_{i=1}^{\rho} \alpha_{31} Y_{t-i} + \sum_{i=1}^{\rho} \alpha_{32} X_{t-i} + \sum_{i=1}^{\rho} \alpha_{33} T_{t-i} + U_{3t}
\end{aligned}
$$

где Y – потребление электроэнергии
X – производство электроэнергии
T – тарифы на электроэнергию
α_{ij} - параметры подлежащие оцениванию

В анализе Variance Decomposition в случае изменения одной из переменных исследуется, на сколько процентов это объясняется самой переменной и на сколько процентов влиянием других переменных. В данном анализе изменения объема потребления электроэнергии в первом периоде объясняются только собой и не зависят от внутренней динамики. В пятом периоде изменения объема потребления электроэнергии на 84,6% исходят от самого себя, на 2,26% от производства электроэнергии и на 13,13% - от тарифов на электроэнергию (Таблица 1).

Таблица 1. Variance Decomposition of Y:

Period	T	X	Y
1	0.000	0.000	100.0
2	0.803	0.118	99.078
3	1.731	1.207	97.061
4	2.517	1.250	96.232
5	13.137	2.263	84.600

Как видно переменная производство электроэнергии незначительно влияет на уровень потребления. Далее проведен анализ импульсных откликов (Рис 1.)

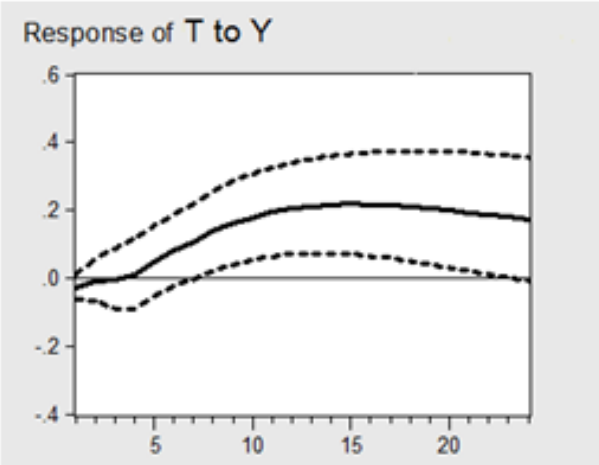


Рис. 1. Анализ импульсных откликов (Impulse Response Analysis)

Анализ функций импульсных откликов определяет изменение переменных в периоды шоков. По результатам модели между переменными обнаружена взаимосвязь: повышение тарифов на электроэнергию после пятого года приводит к значительному снижению потребления электроэнергии, и возвращается на прежний уровень только на двадцать первый год, то есть шестнадцать лет спустя. В данном случае объем потребления электроэнергии зависит не от производства, а от тарифов на электроэнергию.

Оценка статистической значимости осуществлялась на основе P-значений соответствующих t-статистик. Основные проведенные статистические оценки, характеризующие качество модели: коэффициент детерминации $R^2 =$ 0,82, тест Вайта на гомоскедастичность P_{wh} , тест Бреуша-Годфри $P_{BG(2)}$

на отсутствие автокорреляции, при l=2. Приведенные характеристики подтверждают адекватность построенной модели.

3 Выводы

В результате анализа векторной авторегрессии было выявлено, что повышение тарифов на электроэнергию сильно влияет на уровень потребления электроэнергии. В особенности можно предположить, что с момента ввода политики по повышению тарифов, которая планируется в 2014 году, понижение потребления электроэнергии будет достигнуто позже, точнее по полученной модели уже в 2019 году.

Список литературы

1. Федеральная служба государственной статистики РФ. URL: `http://www.gks.ru` (дата обращения 11.12.2013)

A VAR Analysis of Electricity Consumption

Nurgul Mamatova

National Research University Higher School of Economics, Moscow, Russia
nur_gul90@mail.ru

Abstract. This article focuses on the electricity industry and electricity consumption of population and industrial enterprises in Russia is analyzed. Economic factors that affect energy consumption are studied by the Vector Autoregression Analysis (VAR). The relationship between variables is defined by Granger test, and it also considers scenarios of reactions of electricity consumption on exogenous shocks, which are connected with changes in the studied factors.

Keywords: econometric analysis, impulse-response analysis, electricity consumption.

Использование трехмерных анимированных изображений жестов рук для создания анимационной капчи нового типа

Артём Шумилов, Андрей Филиппович

МГТУ им. Н.Э. Баумана, Москва, Россия
{ashumilov,aphilippovich}@it-claim.ru

Аннотация. Статья посвящена созданию нового типа капчи с использованием трехмерных анимированных жестов, которая обладает повышенной сложностью для автоматического распознавания. Описываются этапы создания капчи, анализируются возможные уязвимости и способы улучшения защиты от взлома, рассматриваются перспективы использования. Раскрываются особенности создания 3D-модели руки и жестов, выбранные технологии для представления модели в браузере с помощью кроссбраузерной JavaScript-библиотеки “Three.js”.

Ключевые слова. CAPTCHA, жестомимический интерфейс, распознавание жестов, 3D-моделирование, защита от спама.

1 Введение

Владельцы и администраторы сайтов ежедневно сталкиваются с проблемой спама, целью которого являются реклама, получение доступа к личной информации, создание ссылок для повышения рейтингов в поисковых системах и тому подобное. Очень часто для защиты сайтов от спама и автоматических регистраций используются различные варианты капчи (CAPTCHA) – специальных виджетов, которые предлагают пользователю выполнить простое задание – например, отображают искаженный текст и просят его ввести.

Подобные задания не вызывают трудностей у человека, но оказываются сложны для спам-бота, который после нескольких неудачных попыток прохождения теста переходит к поиску другого сайта с более слабой защитой. Подробнее об изучении эффективности капч представлено в публикациях [1-6].

Из всего многообразия вариантов реализации капчи можно выделить следующие группы: *текстовые, звуковые, математические, логические, образные, интерактивные и анимационные*.

2 Анимационная капча с использованием жестов рук

Основная идея этого вида капчи заключается в том, чтобы показывать пользователю последовательность легко узнаваемых жестов, которые он мог бы воспринимать как последовательность символов или слов. Для обеспечения повышенной сложности автоматического распознавания демонстрируемого жеста предлагается использовать трехмерную интерактивную визуализацию.

В качестве научной базы проекта лежат многолетние исследования в рамках научной школы МГТУ им. Н.Э. Баумана под руководством Ю.Н. Филипповича [7], направленные на создание жестомимического интерфейса [8,9].

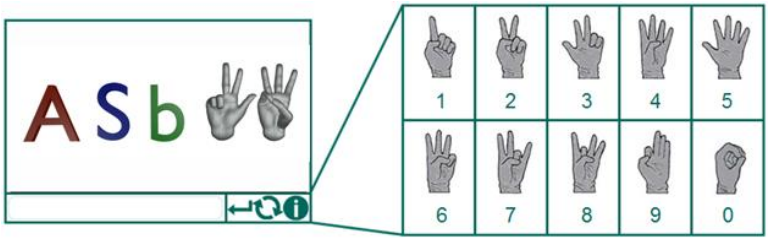


Рис.1. Пример капчи с использованием жестов рук

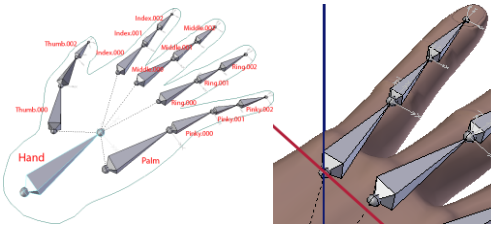


Рис.2. Скелетная модель кисти руки.

Рис.2. Интерфейс управление движением пальцев руки.

Разработка трехмерной модели руки для отображения в браузере. Трехмерную модель руки можно создать в одном из популярных графических редакторов, таких как Autodesk 3d Max, Autodesk Maya, Blender. Для демонстрации трехмерной модели в браузере выбрана кроссбраузерная библиотека Three.js [10]. Для исследования эффективности капчи на первом этапе была разработана модель кисти руки, которая содержит 17 костей.

Разработка алгоритма моделирования движений руки. Трехмерные графические редакторы позволяют создавать также и анимацию. Но эта анимация не предусматривает возможности ее динамического изменения и не может быть использована при создании капчи, так как в этом случае добавление нового жеста в алфавит требует вручную создавать и хранить анимации перехода руки из положения этого жеста в положения всех остальных жестов алфавита.

В связи с этим было принято решение анимировать руку программно, используя прямой доступ к положению костей скелетной модели. Это было сделано с помощью библиотеки Tween.js. Однако, при использовании Tween.js возможны коллизии, когда объекты (в нашем случае, пальцы руки) пересекают друг друга и проходят сквозь друг друга при движении. Поэтому было решено разработать собственный алгоритм моделирования движений и обработки коллизий для создания реалистичной динамической анимации.

Разработка интерфейса создания жестов и перевода текста в жесты. Для того чтобы модель руки могла воспроизводить в виде жестов заданные символы, необходимо создать алфавит жестов. Алфавит должен хранить для каждого символа углы расположения костей относительно друг друга. Для более удобного создания и пополнения алфавита жестов было принято решение разработать графический редактор жестов. С помощью редактора жестов можно также создавать разные алфавиты для различных целевых аудиторий. На рис. 3 показано окно прототипа редактора жестов, позволяющего управлять движением пальцев руки.

Защита капчи от взлома. Относительно небольшое ограниченное количество символов в алфавите может позволить злоумышленнику создать собственную базу данных относительных углов костей, соответствующих символам алфавита. Поэтому имеет смысл рассмотреть возможность преобразования сформированной анимации в формат GIF для демонстрации его пользователю в качестве капчи. В дальнейшем также предполагается отработать и другие методы защиты – повышение полигональности модели, изменение текстур, добавление шумов и т.д.

Понимание жестов человеком. В настоящее время количество общепринятых или интуитивно понятных жестов руки достаточно невелико, поэтому использование только таких жестов диктует достаточно небольшой размер алфавита. Алфавит может быть расширен в случае использования на специализированных сайтах, предназначенных для людей, владеющих каким-либо жестовым языком, например, на сайтах, предназначенных для подводников, музыкантов, спортсменов и т.д.

Тем не менее, подобная капча применима на любых сайтах, так как в задании теста жесты могут быть показаны несколькими моделями рук, а также скомбинированы с вращающимися моделями букв латинского или кириллического алфавита. Таким образом, может быть зашифрован цифробуквенный код, который применяется в большинстве современных капч. Кроме того, жестовая капча может содержать несколько моделей рук и предлагать пользователю выбрать определенный символ на основании того, какая именно в данный момент используется модель.

3 Заключение

У предлагаемого варианта капчи хорошие перспективы использования. Во-первых, пользователю предлагается качественное, легкое для восприятия трех-

мерное изображение знакомого жеста, распознавание которого не вызовет у него затруднений и не потребует длительного времени. Во-вторых, этот вариант реализации капчи относится к анимационной группе, которая считается самой сложной для автоматического распознавания.

У проекта также могут быть перспективы развития в части создания тематических (настраиваемых администраторами) серий жестов для различных профессиональных и социальных групп, в том числе и с ограниченными возможностями. Развитие мобильных технологий позволяет потенциально использовать технические возможности устройств для повторения жестов вместо набора, а также применения соответствующих алгоритмов для задач идентификации и авторизации пользователей.

Жестовая капча может использоваться в качестве альтернативного варианта вместо аудио тестирования, что позволит людям с проблемами восприятия звуков или не имеющим аудио оборудования пройти тест.

Список литературы

1. Converse, T.: CAPTCHA Generation as a Web Service. In: Baird, H.S., Lopresti, D.P. (eds.) HIP 2005. LNCS, vol. 3517, pp. 82-96. Springer, Heidelberg (2005)
2. Moy, G., Jones, N., Harkless, C., Potter, R.: Distortion Estimation Techniques in Solving Visual CAPTCHAs. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), vol. 2, pp. 23-28 (2004)
3. Vicarious AI passes first Turing Test: CAPTCHA [Электронный ресурс]. URL: http://news.vicarious.com/post/65316134613/vicarious-ai-passes-first-turing-test-captcha
4. Анимационная CAPTCHA легче для людей и тяжелее для ботов [Электронный ресурс]. URL: http://www.aiportal.ru/news/animated-captcha.html
5. David Bushell. In Search Of The Perfect CAPTCHA. [Электронный ресурс]. URL: http://coding.smashingmagazine.com/2011/03/04/in-search-of-the-perfect-captcha.
6. Jeff Atwood. CAPTCHA Effectiveness. [Электронный ресурс]. URL: http://www.codinghorror.com/blog/2006/10/captcha-effectiveness.html.
7. Филиппович А.Ю. Научно-образовательный кластер в интернете // Качество образования, сентябрь 2012 – С. 40-45
8. Филиппович Ю.Н. Компьютерные средства поддержки коммуникативного взаимодействия людей с ограниченными слуховыми возможностями. Proceedings of 10th International Congress of the international society of applied Psycholinguistics “Challenges of information Society and applied psycholinguistics”., RUDN-Institute of Linguistics RAN-MIL, Москва, 2013, С. 254
9. Филиппович Ю.Н., Зеленцов И.А Распознавание скорописи XVII в.// Проблемы полиграфии и издательского дела. – 2011. – № 3. – С. 87-97.
10. [Электронный ресурс]. URL: http://threejs.org/.

Using 3D Animated Hand Gestures to Create a New Type of CAPTCHA

Artem Shumilov, Andrew Philippovich

Bauman Moscow State Technical University, Moscow, Russia
{ashumilov,aphilippovich}@it-claim.ru

Abstract. Website owners and administrators have to deal with the problem of spam every day. To protect their websites from spam webmasters use CAPTCHA – special tests created to tell computers and humans apart. This article focuses on one of the most difficult for automatic recognition type of CAPTCHA using three-dimensional animated images hand gestures.

Keywords: CAPTCHA, gesture and mimic interface, gesture recognition, 3D modeling, spam protection.

Автоматическая расстановка рейтинга музыкальным произведениям на основе неявных оценок

Сергей Смагин

Тульский государственный университет, Тула, Россия
smaginsergey1310@gmail.com

Аннотация В настоящее время людям приходится взаимодействовать с огромным количеством различных данных. Помочь ему в этом призваны рекомендательные системы. В частности, у слушателя музыки на электронных устройствах обычно достаточно большой список воспроизведения, и не каждая композиция в таком списке ему может одинаково нравиться. В данной статье описан подход к вычислению рейтинга музыкальных композиций на основе неявных оценок.

Ключевые слова: рекомендательные системы, рейтинг музыки.

1 Введение

Программными музыкальными проигрывателями пользуется большое число человек. У многих слушателей большой список воспроизведения, который часто состоит из песен, которые пользователь не хотел бы слушать по той или иной причине (например, они надоели или их скачали вместе с другими песнями того же исполнителя и она оказалась неудачной). Одним из решений такой проблемы является методичное переслушивание всего списка воспроизведения и удаление нежелательных к прослушиванию треков. Недостатки этого подхода очевидны:

- Не каждый захочет уделить несколько часов (а в худшем случае — дней) на приведение коллекции в порядок.
- Будут появляться новые песни, старые будут надоедать и через некоторое время список воспроизведения снова будет наполнен композициями, которые не хочется слушать.

Другим решением этой проблемы может быть рекомендательная система. Пользователь сам оценит некоторые треки, чтобы система проигрывала их чаще других.

Как известно, рекомендательные системы бывают построены на явных или неявных оценках[1]. Пример рекомендательной системы с явными оценками — применяемая во многих проигрывателях схема со “звездами” вместо

[1] https://ru.wikipedia.org/wiki/Рекомендательная_система/Методика

оценок. Но для большинства людей такая система также неэффективна — расставлять оценки каждый раз нужно вручную, а это ненамного лучше ручного удаления композиций.

Рекомендательная система на основе неявных оценок в этом случае выглядит гораздо предпочтительнее, поскольку прослушивание музыки в большинстве случаев — фоновая задача. Неявными оценками в данном случае будут действия пользователя музыкального проигрывателя.

2 Существующие решения

В Amarok[2] и Clementine[3] используется следующее решение. Добавляемая в коллекцию песня имеет средний рейтинг и далее, в зависимости от доли прослушанной композиции, рейтинг увеличивается или уменьшается.

Многие программы-проигрыватели базовой поставке или в виде расширений имеют систему оценки композиций с применением «звезд».

В Интернете существует сразу несколько известных сервисов для оценки музыки (самый известный, должно быть, Last.fm[4]), для оффлайновых проигрывателей нет широко известной рекомендательной системы.

3 Предлагаемое решение

Для решения задачи был придуман алгоритм, основанный на следующих поведенческих признаках:

– Много раз прослушанная песня нравится меньше услышанной впервые.
– Песни, которые не нравятся, переключают.

Тогда, во-первых, рейтинг должен уменьшаться с количеством прослушиваний. В разработанном алгоритме применена следующая формула:

$$r = \frac{1}{lg(10 + i \cdot lc)}, \tag{1}$$

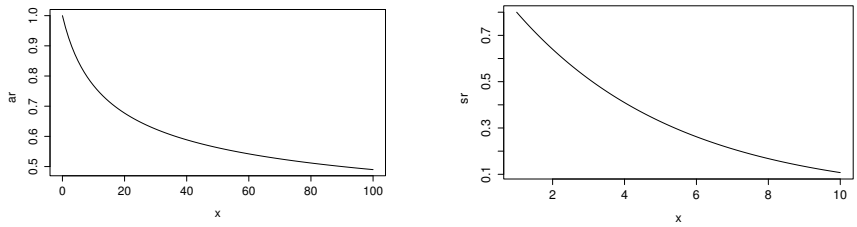
где i — коэффициент влияния возраста, показывает, с какой скоростью песни устаревают (по умолчанию рейтинг уменьшается вполовину после 90 прослушиваний) и lc — количество прослушиваний.

Во-вторых, рейтинг песни должен быстро (гораздо быстрее, чем при прослушивании) уменьшаться при переключении. Была применена формула:

$$nr = or \cdot (1 - si \cdot (1 - sr)), \tag{2}$$

где nr, or — новое и старое значения рейтинга песни, si — коэффициент влияния переключения песни (по умолчанию рейтинг песни может максимально уменьшиться на 0.2 от прошлого значения), sr — прослушанная доля песни.

[2] `http://amarok.kde.org/`
[3] `http://www.clementine-player.org/`
[4] `http://www.lastfm.ru/`



(a) Снижение рейтинга песни в зависимости от количества прослушиваний

(b) Снижение рейтинга песни в зависимости от количества пропусков

Рис. 1: Графики уменьшения рейтинга песни

Формулы (1) и (2) формируют рейтинги композиции (максимальное значение по каждому критерию — единица). Для получения общего рейтинга нужно перемножить значения критериев. Начальное значение по каждому следует взять максимальным, поскольку новые песни хочется прослушать сразу после добавления. Достоинства такого подхода:

- Конфигурируемость — пользователь может задать соответствующие коэффициенты влияния по своему вкусу.
- Неявные оценки — пользователю не нужно делать ничего специального для того, чтобы система работала.

Данный алгоритм реализован в программе autorating[5] как клиент для mpd[6] — системной службы для проигрывания музыки. Данная реализация оказалась очень компактной и простой за счет клиент-серверной архитектуры mpd.

4 Оценка качества рекомендаций

Для оценки качества рекомендаций поставим следующий эксперимент. Возьмем несколько песен разной длины. Пусть одна из них условно нравится, другие – нет. Ненравящиеся песни будем пропускать на первых секундах. Будем проигрывать песни 5 кругов. Перед началом проигрывания каждая песня имеет рейтинг 255 (наивысший возможный).

Как видно, песня, условно нравящаяся, по рейтингу далеко опережает те, которые не нравятся, см. таблицу 1. Кроме того, рейтинг последних крайне мало зависит от длины, если треки пропускать на первых секундах (как это обычно делается с ненравящимися композициями).

[5] `https://github.com/s-mage/autorating`
[6] `http://www.musicpd.org/`

Таблица 1: Сравнение рейтинга песен после окончания эксперимента

Название	**Длина, с**	**Нравится?**	**Рейтинг**
mix 25	3540	нет	70
striken	245	да	216
remember the name	220	нет	71
hybris	206	нет	71
i miss you	366	нет	70
а ну отдай мой каменный топор	120	нет	72

5 Заключение

Представлен алгоритм автоматического оценивания музыкальных произведений на основе неявных оценок. Данный алгоритм относительно прост и основывается на поведенческих признаках, которые подходят большинству людей. При испытании программы, реализующей данный алгоритм, была подтверждена ее способность к решению поставленной задачи и эффективность.

Направления дальнейшей работы перечислены ниже.

- Больше поведенческих признаков. Например, переключение после паузы может не учитываться (по личному опыту, не всегда хочется слушать даже хорошую песню не с начала).
- Добавление возможности пользователю явно выразить отношение к песне и учитывать это в общем рейтинге.
- Подстройка под настроение. В зависимости от настроения человек может быть расположен слушать разные песни.

Список литературы

1. Kordumova, Suzana et al. Personalized implicit learning in a music recommender system. User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg, 2010. 351-362.
2. Kim, Hyun-Jun, and Young Sang Choi. EmoSens: Affective entity scoring, a novel service recommendation framework for mobile platform. Workshop on personalization in mobile application of the 5th international conference on recommender system. 2011.
3. The International Society for Music Information Retrieval, `http://www.ismir.net`.
4. Zaharchuk, Vasily et al. A new recommender system for the interactive radio network fmhost. Proceedings of the international workshop on experimental economics and machine learning (EEML). 2012.

Automatic Music Rating Based on Implicit Assessments

Sergey Smagin

Tula State University, Tula, Russia
smaginsergey1310@gmail.com

Abstract. Nowadays people have to interact with huge amount of data. Recommender systems are all about help them with it. Particulaly, music listener often has large playlist and not each track he likes the same. This article describes a way to evaluate music rating based on implicit assessments.

Keywords: music rating, recommender systems.

Алгоритм семантического поиска в больших текстовых коллекциях

Виталий Савченко

АлтГТУ им. И. И. Ползунова, Барнаул, Россия
64svv@rambler.ru

Аннотация В статье рассматривается метод семантического поиска для многопоточной обработки текстов большого объема. Поисковый запрос и обрабатываемый текст преобразуются в графы семантических связей. Предлагается алгоритм вычисления коэффициента соответствия семантических графов. Приводятся оценки времени обработки.

Ключевые слова: семантический анализатор, граф, справочник, вес, тип семантической связи.

1 Общие сведения

В наше время, в условиях большого и стремительно растущего объема информации, актуальна задача поиска в больших текстовых коллекциях [1]. Одним из вариантов поиска является семантический поиск, т.е. поиск с точки зрения содержащейся в тексте информации [2,3,4]. Среди наиболее популярных систем семантического поиска можно выделить Google, SearchMonkey, Freebase и AskNet. Однако они имеют определенные недостатки, такие как: применение семантики лишь для незначительного улучшения результатов поиска, ограничение на длину запроса, снижение качества поиска с увеличением поискового запроса. Кроме того большинство из таких поисковых систем работают только с английским языком.

Учитывая сложность семантического поиска необходимо применять методы, основанные на имеющихся в системе знаниях о предметной области.

2 Семантический поиск

В данной работе представлен результат разработки системы семантического поиска для больших текстовых коллекций на русском языке. Ключевой особенностью полученной системы - является снятие ограничений на величину поискового запроса и многопоточная обработка текстовой коллекции.

Исходными данными для поиска являются текстовые коллекции и запрос пользователя, который представляет собой текстовую коллекцию. Исходя из

предположения, что большая текстовая коллекция в общем случае неоднородна и с точки зрения поиска интересна ее определенная часть, то текст нужно разделить на определенные участки - страницы, абзацы или наборы из нескольких предложений. Такие фрагменты будем называть «окнами».

Для каждого окна запроса и поисковых коллекций строится граф семантических связей, назовем его «семантический граф». Семантический граф представляет собой направленный граф, вершинами которого являются слова русского языка, представленные в нормальной форме, а ребра характеризуются весом и типом семантической связи. Направление ребра зависит от типа семантической связи, например, отношение объект - действие, объект - свойство, действие - время.

Для построения семантического графа каждое предложение из окна коллекции обрабатывается семантическим анализатором. В данной работе используется семантический анализатор RML[1].

Предложения окна обрабатываются последовательно. На каждой итерации семантический граф предыдущей итерации объединяется с графом G_{new} обрабатываемого предложения. Веса ребер семантического графа G_{new} равны 1. После объединения у графа G_{i+1} все веса ребер умножаются на коэффициент затухания η.

$$G_{i+1} = (G_i + G_{new}) * \eta \tag{1}$$

После этого результирующий семантический граф используется для следующей итерации. Затем из графа удаляются ребра с весом меньше δ. Уменьшение веса ребер на заданный процент, аналогичное испарению феромона в «муравьином алгоритме» [5], сделано для ослабления воздействие предшествующих семантических зависимостей между вершинами графа.

Далее необходимо подсчитать величину коэффициента соответствия семантического графа запроса и семантического графа окна. Простой поиск наибольшего общего подграфа, даже с учетом совпадения не только вершин, но и типов ребер, не приведет к цели. Во-первых, по причине NP- полноты данной задачи. Во-вторых, один и тот же смысл содержится в текстах разного стилевого оформления, например, содержит обобщающие сведения или только частичную информацию. К хордовым, обитающим в тайге, в том числе относятся и зайцы тайги. Очевидно, улавливается связь: хордовые $\rightarrow$ зайцы. Однако между хордовыми и зайцами не должно быть полного отождествления, т. к. хордовые – это не только зайцы.

Для поиска связанных по смыслу слов был использован словарь, в котором представлен перечень слов в нормальной форме [6]. Каждому слову сопоставлен набор слов, связанных с ним ассоциативной, синонимичной и т. д. связью. Таким образом, словарь представляет собой направленный граф $G_{word} = (V_{word}, U_{word})$, где вершины V_{word} - это слова в нормальной форме, а ребра U_{word} имеют действительные весовые коэффициенты от 0 до 1. Назовем граф G_{word} графом справочника.

[1] `http://www.aot.ru`

За коэффициент связанности слов a_k и a_m - вершин семантического графа запроса $G_{request}$ и семантического графа окна коллекции G_{text} возьмем произведение весов от таких же слов до общего предка в графе справочника. При совпадении слов данный коэффициент будет равен 1, иначе будет принадлежать промежутку [0;1].

Замечание: Граф справочника является упрощенной моделью знаний о реальном мире, а общий предок слов a_k и a_m в этом графе - это некоторое обобщение соответствующих понятий. Нет смысла искать общего предка слов во всем графе справочника. Следовательно, нас интересует некое ξ окружение искомых слов. В противном случае считаем, что слова никак не связаны по смыслу. Фрагмент графа G_{word} представлен на рис. 1.

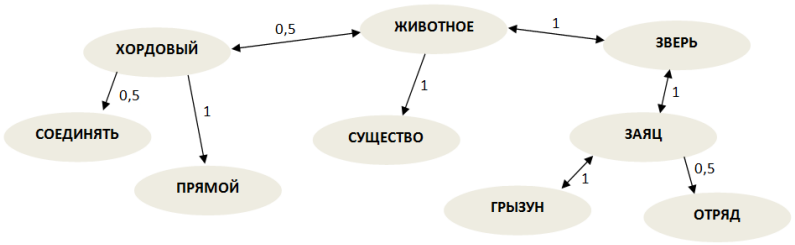


Рис. 1. Граф справочника

Далее ищем наилучшее совпадение семантического графа запроса с графом окна. Коэффициент соответствия рассчитываем по формуле 2:

$$S = \sum_{k=1}^{n} D1_k * D2_k * L1_k * L2_k, \tag{2}$$

где n – количество совпавших ребер графа запроса и окна, k – индекс соответствующего ребра, $D1_k$, $D2_k$ – произведение весов ребер в графе справочника от слова запроса и слова окна до общего предка соответственно, $L1_k$ – вес ребра k семантического графа запроса, $L2_k$ – вес ребра k семантического графа окна. Так как вариантов совпадения графов много – нас интересует максимальное значение коэффициента соответствия. Определив максимальное значение по всем окнам текстовой коллекции - получим общее значение – коэффициент соответствия запроса и текстовой коллекции.

3 Тесты и результаты

Оценка полученной системы является экспертной. Для поиска были отобраны текстовые коллекции большого объема удовлетворяющие одному и тому же запросу к поисковой системе google.ru. В зависимости от настроек

системы были получены различные значения коэффициента соответствия между поисковым запросом и коллекцией. Однако, для коллекций по содержанию которых строился запрос или коллекций аналогичного содержания значение коэффициента минимум на порядок превосходило значение коэффициента других коллекции, не похожих по содержанию.

Временные затраты на обработку коллекции в зависимости от количества предложений в запросе и окне коллекции, а так же относительные временные затраты при многопоточной обработке представлены на рис. 2.

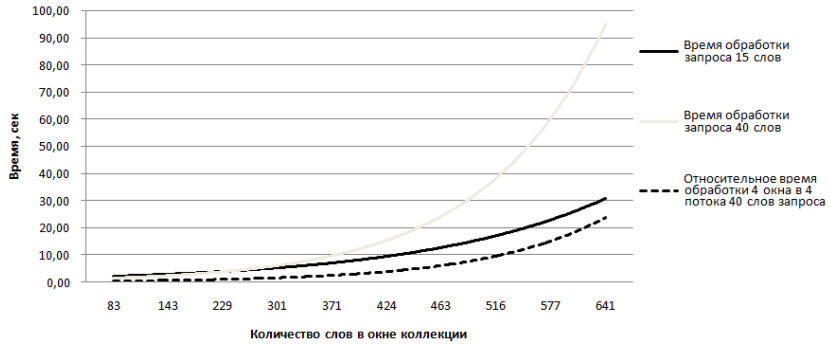


Рис. 2. Временные затраты

Основным минусом текущей реализации алгоритма является значительное увеличение времени обработки с ростом размеров окна и запроса. Плюсом является то, что текстовая коллекция и запрос могут быть разделены на окна оптимального размера с точки зрения времени обработки. Кроме того обработка окон текстовой коллекции, в данном случае, может выполняться параллельно, что значительно повышает скорость выполнения на многопоточных и многопроцессорных системах.

Список литературы

1. *Hannah Bast, Marjan Celikik* Efficient Fuzzy Search in Large Text Collections // ACM Transactions on Information Systems, 2010.
2. *Mathieu d'Aquin, Enrico Motta* Watson, more than a Semantic Web search engine // IOS Press Amsterdam, 2011.
3. *K Elbedweihy, S N Wrigley, F Ciravegna, D Reinhard, A Bernstein* Evaluating Semantic Search Systems to Identify Future Directions of Research // Second International Workshop on Evaluation of Semantic Technologies, page 25-36, 2012.
4. *G. Tsoumakas, M. Laliotis, N. Markantonatos, I. Vlahavas* Large-Scale Semantic Indexing of Biomedical Publications at BioASQ // BioASQ Workshop, 2013.
5. *Штовба С. Д.* Муравьиные алгоритмы // Exponenta Pro. Математика в приложениях, №4, с.70-75, 2003

6. *Крайванова В.А., Кротова А.О., Крючкова Е.Н.* Построение взвешенного лексикона на основе лингвистических словарей // Материалы Всероссийской конференции с международным участием ЗОНТ-2011, Т.2, Новосибирск, 2011.

Semantic Search Algorithms in Large Text Collections

Vitaliy V. Savchenko

Altai State Technical University, Barnaul, Russia
64svv@rambler.ru

Abstract. This article describes a method of semantic search based on the text processing of large volume. Search requests and processing text from analyzed collection is transformed into a graph of semantic relationships, the comparison of which allows us to define a measure of semantic similarity of compared texts. An algorithm is proposed to calculate the coefficient of semantic graphs concordance. Estimates of the processing time are also given.

Keywords: semantic analyzer, graph, directory, weight, type of semantic communication.

Организация массового свободного ассоциативного эксперимента в сети Интернет при помощи модуля CAPTCHA

Дмитрий Лахвич

МГТУ им. Н.Э. Баумана, Москва, Россия
dlakhvich@it-claim.ru

Аннотация. Статья посвящена особенностям проведения свободного ассоциативного эксперимента. Рассмотрены основные факторы влияющие на качество результатов эксперимента. Продемонстрован новый способ проведения эксперимента при помощи модуля CAPTCHA, позволяющий сократить временные расходы на проведение эксперимента.

Ключевые слова: свободный ассоциативный эксперимент, CAPTCHA, блог, маркетинговые исследования.

1 Введение

Как показывает практика диапазон применения свободного ассоциативного эксперимента (САЭ) для решения различных задач весьма широк [1]. Ввиду своей относительной простоты он прекрасно подходит для решения теоретических и практических задач в области психолингвистики, психиатрии, маркетинге, оценки трендов и т.д. Само задание достаточно простое и не вызывает затруднений у респондента. При проведении САЭ важными факторами являются [2–4]:

- Формат проведения эксперимента. Эксперимент может проводится устно или письменно, групповое или индивидуальное предъявление списка слов.
- Уровень квалификации экспериментатора.
- Место проведения, время года и суток, окружающая температура и т.д.
- Технология обработки полученного ассоциативного материала.
- Индивидуальные характеристики респондента.
- Достоверность полученных данных.

Все эти моменты обуславливают собой систему факторов, оказывающую сильное воздействие на достоверность полученных ассоциативных данных, а следовательно — выводов, сделанных на основе них. Некоторые этих факторов хорошо изучены, некоторые – нет [5].

2 Автоматизированные системы проведения САЭ

Взрывной рост информационных технологий подогрел интерес исследователей к САЭ, а также сделал возможным проведения массового САЭ с помощью специализированых настольных (desktop) приложений и web-приложений, которые позволяют значительно расширить географию эксперимента при снижении временных и человеческих ресурсов [3]. Автоматизированные средства также позволили упростить последующую обработку, хранение и систематизацию полученных данных, а также, самое главное, позволили выделить новые закономерности [1].

В качестве основных отрицательных факторов большинства информационных систем (ИС) можно выделить следующие факторы:

- в большинстве случаев отсутствует возможность проконтролировать САЭ;
- проблематичность опроса неподготовленного респондента;
- Отсутствие эргономичности ИС;
- перегруженность пользователя информацией;
- респондент быстро теряет интерес к заполнению больших форм ответов.

С точки зрения автора, последний из представленных факторов является важнейшим при проведении САЭ через сеть Интернет. Чем сложнее и искусственнее экспериментальная ситуация, тем менее информативны оказываются получаемые экспериментальные данные. Испытуемый должен быть вовлечен в эксперимент ровно в той степени в которой необходимо. Так, например, большое количество слов-стимулов может вызвать у испытуемого отторжение, наблюдается потеря внимания, усталость, снижается интерес к эксперименту в целом. Это приводит к резкому возрастанию экстрасигнальных реакций и отказов от эксперимента вообще. Интернет обостряет требование к минимизации получения слов-реакций за такт проведения эксперимента, пользователь Интернета привыкший к интерактивности и возможности “случайно” потерять данные, более охотно идёт на заполнение множества мелких анкет, нежели одной большой.

3 Проведение САЭ с использованием модуля CAPTCHA

Для того чтобы снизить навязчивость эксперимента предлагается использовать модифицированный модуль CAPTCHA (от англ. Completely Automated Public Turing test to tell Computers and Humans Apart — полностью автоматизированный публичный тест Тьюринга для различения компьютеров и людей) в блог/новостных движках сайтов-партнёрах. В рамках предлагаемой работы модификации подвергается классический вид, только теперь помимо вывода картинки, происходит вывод от одного до трёх слов-стимулов с дополнительной просьбой заполнить поля для слов-реакций.

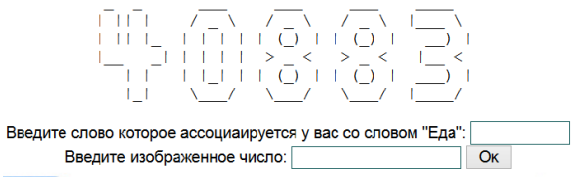


Рис. 1. Пример реализации

Запрос CAPTCHA происходит при попытке пользователя прокомментировать некоторую новость или публикацию. Таким образом появляется возможность провести САЭ в более игровой форме, в некотором добровольно-принудительном для пользователя порядке, при этом большинство пользователей, уже привыкших к CAPTCHA не будут испытывать больших затруднений, а сложность прохождения для “ботов-спамеров” будет значительно увеличена. В будущем собрав достаточно данных можно будет отказаться от вывода самой изображения, оставив только требования ввода слов реакций.

Данное изменение в принцип проведения автоматизированного САЭ позволит снизить порог вхождения для пользователя, а, следовательно, позволит провести по-настоящему массовый свободный ассоциативный эксперимент.

Привязка слов-реакций на слова-стимулы к реальному аккаунту пользователя (а не специально зарегистрированному для проведения эксперимента) позволяет проводить различные исследования различных групп пользователей (географическое, социальное, культурное деление групп пользователей). Также это позволит практически исключить “сознательные шумы” генерируемы недобросовестными пользователями.

Так как попытка ответа на какую-либо публикацию непроизвольно инициирует запуск САЭ, мы можем исследовать полученный материал в разрезе реакций на различные явления жизни, будь то новость о свадьбе одной из европейских принцесс, так и новостей о катастрофах унесших сотни человеческих жизней.

4 Техническая реализация

В качестве платформы для реализации поставленной задачи была выбрана блог-платформа WordPress, за то, что обладает большим количество плагинов под свободными лицензиями.

В качестве языков программирования: JavaScript для создания “тонкого клиента” и сбора дополнительной статистики о поведении пользователя, и PHP для реализации серверной части ответственной за сбор данных и генерацию допол-

нительных форм. В качестве базы используется MySQL.

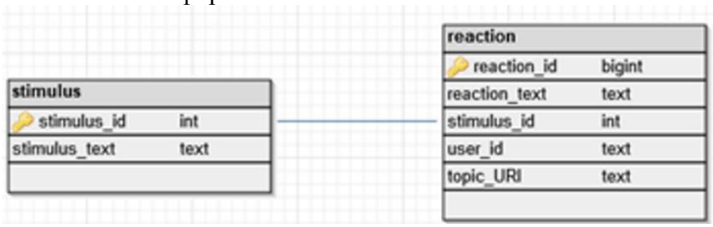


Рис. 2. Необходимая для добавления в базу данных схема отношений

Основной задачей, которую предстоит решить, – поиск сайтов-партнеров, которые будут готовы поделиться правом собирать статистику. Ставится цель найти сайты различной тематики для повышения репрезентативности получаемых данных.

Список литературы

1. Филиппович Ю.Н., Филиппович А.Ю. Динамическая инфокогнитивная модель вербального сознания // «Нейрокомпьютеры: разработка, применение», №1, 2013 г. - С.13-22.
2. Филиппович А.Ю. АСНИ ассоциативных экспериментов // Вопросы психолингвистики. 2007. № 6. - С. 143-153.
3. Филиппович А.Ю., Кирнарский А.Б. Проведение интерактивных лингвистических ассоциативных экспериментов в сети Интернет // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 8 / Сост. и ред. Ю.Н. Филиппович. — М.: НОК «CLAIM», 2006. - С. 96-106
4. Черкасова Г.А. Исследование динамики ассоциативно-вербальной модели языкового сознания русских // Вопросы психолингвистики, 6, 2007. М, 2008. с. 105-122.
5. Горошко Е.И. Интегративная модель свободного ассоциативного эксперимента. – Харьков; М.: Изд. группа «РА-Каравелла», 2001. – 320 с

Using CAPTCHA in a Massive Free Association Experiment on the Internet

Dmitry Lakhvich

Bauman Moscow State Technical University, Moscow, Russia
dlakhvich@it-claim.ru

Abstract. The paper is focused on organizing a massive free association experiment. The primary quality affecting factors have been analyzed. The present CAPTCHA-based approach for performing an experiment allows one to reduce the time spent on performing an experiment.

Keywords: free association experiment, CAPTCHA, blog, market research.

Определение характеристик городов, влияющих на тональность отзывов, на основе анализа социальной сети Twitter

Александр Зырянов, Никита Путинцев

Exposoft, Новосибирск, Россия
{alexander.zyryanov44,putintsevnikita}@gmail.com

Аннотация Статья посвящена анализу сообщений в социальной сети Twitter. В ходе работы устанавливается, какие характеристики городов России влияют на тональность сообщений, посвященных тому или иному городу, другими словами, от каких характеристик зависит отношение людей к городу.

Ключевые слова: тональность текста, FRiS, машинное обучение, кластеризация.

1 Введение

Сейчас многие населенные пункты в России теряют свое население в пользу более крупных и оживленных городов, о чем можно судить на основании данных Росстата[1] и Госкомстата[2]. Уезжают в основном молодые и перспективные люди, при этом обратный приток населения незначителен. Это ведет к уменьшению уровня производства местных предприятий вследствие недостатка кадров, к уменьшению качества образования в местных школах и университетах, к ухудшению экономического и социального состояния городов в целом. Эта проблема становится все более актуальной и ее решение – совсем нелегкая задача, требующая глубокого понимания причин, которые её формируют. В связи с этим возникает интерес попытаться выявить основные движущие факторы этой проблемы при помощи анализа текстовых сообщений в социальных сетях.

В социальных сетях люди охотно высказывают свое мнение по любому вопросу. Причем, в отличие от соцопросов, где люди часто отвечают неохотно, не задумываясь, так что их мнение искажено или не соответствует действительности, в сетях высказывания зачастую сформированы настоящими мыслями людей. Кроме того, социальные сети могут предоставить миллионы сообщений для обработки практически даром, тогда как для проведения опроса такого же объема потребуются значительные затраты времени и ресурсов. Извлекая из этих сообщений мнения о различных городах, мы можем выяснить насколько хорошо или плохо люди к ним относятся.

[1] http://www.gks.ru/free_doc/doc_2013/bul_dr/mun_obr2013.rar
[2] http://www.gks.ru/free_doc/new_site/perepis2010/croc/perepis_itogi1612.htm

2 Постановка задачи

В основе нашего исследования лежит предположение о том, что тональность сообщений об определенном городе и тональность сообщений, сделанных из этого города в различных социальных сетях, может зависеть от его социальных, экономических и географических характеристик. Предполагается исследовать влияние таких характеристик, как плотность населения, климат, средний уровень заработной платы, возрастной состав, половой состав, наличие крупных торговых центров, парков и зон отдыха.

Для решения поставленной задачи в первую очередь необходимо набрать базу сообщений, которые относятся к определенным городам России или были в них созданы. Для этого нужно, во-первых, найти источник информации и, во-вторых, отфильтровать нужные для исследования сообщения. Далее необходимо определить тональность собранных высказываний.

Следующим шагом необходимо набрать информацию о выбранных характеристиках городов России и привести ее к удобному для работы виду.

Наконец, планируется выделить те характеристики городов, которые оказывают наибольшее влияние на тональность сообщений. Опционально города планируется разбить на таксоны и определить, к какому типу городов люди относятся лучше всего.

3 Аналогичные работы

Идея использовать Twitter для анализа мнений людей по различным вопросам возникла довольно давно [2]. Существуют похожие исследования, в которых тональность сообщений используется для предсказания каких-либо событий [1]. Так же уже разработано и опробовано большое количество различных методов анализа тональности сообщений, как использующих словарь эмотивной лексики [4], так и обучающихся на выборке [3], [5] . Эти методы оказались достаточно эффективными и подходят для поставленной в данной работе задачи.

4 Предполагаемое решение

В качестве источника сообщений решено использовать Twitter, так как он имеет широкую и разнообразную аудиторию, и содержит огромное количество сообщений, которое растет с каждым днем. Кроме того данная сеть предоставляет API для работы с потоком новых сообщений и данные в качестве грантов[3]. Для отбора сообщений о городах используется, во-первых, словарь их полных и сокращенных названий в различных морфологических формах, во-вторых геолокация. Для фильтрации спама на обучающей

[3] `https://blog.twitter.com/2014/introducing-twitter-data-grants`

выборке тренируется наивный классификатор Байеса. Сообщения подвергаются предварительной обработке, которая включает нормализацию сообщений, осуществляемую при помощи Pymorphy[4], и исключение стоп-слов. Стоп-слова планируется убрать автоматически, используя индекс TF-IDF на всей коллекции собранных сообщений из Twitter[6],а так же находящиеся в открытом доступе словари. Так же планируется заменить все эмотиконы на специальные слова, соответствующие их тональности.

Для оценки тональности полученных сообщений выбраны наивный классификатор Байеса из-за его простоты и эффективности, а так же метод опорных векторов из-за его точности [5]

Отдельный интерес представляет использование алгоритма классификации FRiS Stolp [7]. Интерес обусловлен желанием проверить пригодность данного алгоритма для решения задач анализа текстов.

Информацию о городах планируется собрать в полуавтоматическом режиме, используя интернет ресурсы, в частности, Wikipedia. Для кластеризации городов будет использован алгоритм FRiS Tax [7].

Для определения наиболее значимых признаков предлагается использовать способность алгоритма Random Forest определять важность используемых признаков [8]. Достаточно просто обучить алгоритм на таблице объект-свойство всех городов с целевым признаком тональности, который высчитывается как сумма тональностей всех сообщений, относящихся к данному городу.

5 Заключение

В работе обозначена задача выявления характеристики городов, которые оказывают влияние на тональность сообщений в социальных сетях. Так же представлено предполагаемое решение этой задачи, основанное на анализе тональности сообщений социальной сети Twitter методами машинного обучения.

Список литературы

1. *Bollen, J., Maon, H., Zeng, H.* Twitter mood predicts the stock market // Journal of Computational Science. Март 2011. № 1(2). С. 1–8.
2. *Pak, A., Paroubek, P.* Twitter as a Corpus for Sentiment Analysis and Opinion Mining // Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010
3. *Kouloumpis, E., Wilson, T., Moore, J.* Twitter sentiment analysis: The good the bad and the omg! // The AAAI Press, 2011
4. *Клековкина, М. В., Котельников, Е. В.* Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики /в сб. Труды XIV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». С. 118–123. — Переславль-Залесский: изд-во «Университет города Переславль», 2012.

[4] `https://pythonhosted.org/pymorphy/`

5. *Клековкина, М. В., Котельников, Е. В.* Автоматический анализ текстов на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). – Вып. 11 (18). – М. : Изд-во РГГУ, 2012.
6. *Ramos, J.* Using TF-IDF to Determine Word Relevance in Document Queries // The First Instructional Conference on Machine Learnin, 2003.
7. *Borisova, I. A., Dyubanov, V. V., Kutnenko, O. A., Zagoruiko, N. G.* Use of the FRiS-Function for Taxonomy, Attribute Selection and Decision Rule Construction /в сб «Lecture Notes in Computer Science» С. 256–270. — Berlin: Springer Berlin Heidelberg, 2011.
8. *Breiman, L.* Random Forests // Machine Learning, 2001. Т. 45. № 1. С. 5–32.

Determining Which Cities' Features Affect the Opinions' Sentiments on Twitter

Alexander Zyryanov, Nikita Putintsev

Exposoft, Novosibirsk, Russia
{alexander.zyryanov44,putintsevnikita}@gmail.com

Abstract. The paper is devoted to analysis of messages in the Twitter social network. The present study is focused on which Russian cities' features do affect the opinions' sentiments expressed by people.

Keywords: text sentiment, FRiS, machine learning, clustering.

Общественное мнение онлайн: сравнение структуры и тематики постов «обычных» и «популярных» блогеров Живого Журнала

Светлана Алексеева, Олеся Кольцова, Сергей Кольцов

НИУ ВШЭ, Санкт-Петербург, Россия
{salexeeva,ekoltsova,skoltsov}@hse.ru

Аннотация. Статья посвящена сравнению тематической структуры и основных статистических параметров постов «обычных» и «популярных» блогеров Живого Журнала. Исследование показало существенное тематическое сходство обеих выборок, была опровергнута гипотеза о большем интересе «топовых» блогеров к социально-политическим темам по сравнению с обычными блогерами. Различие между двумя группами заключается в меньшей активности и большей зашумленности данных среди «обычных» пользователей.

Ключевые слова: тематическое моделирование; LDA; Живой Журнал; общественное мнение.

Введение

В настоящее время в сообществе интернет-профессионалов установилось представление, что блогосфера, наряду с другим пользовательским контентом, является важным источником общественного мнения интернет-активной части населения [1, 2]. В русскоязычном сегменте большая часть общественно значимых дискуссий сосредоточена на платформе Живого Журнала, поэтому именно этот ресурс выбран предметом исследования [3].

Перед исследователем само-сгенерированного общественного мнения в Живом Журнале стоит ряд методологических вопросов, которые необходимо решить перед проведением собственно социологического исследования. В частности, посты каких блогеров — «обычных» или «популярных» — выбирать для анализа? Какое количество текстов выбирать для анализа, если известно, что за неделю на страницах первых 2000 «топовых» аккаунтов Живого Журнала появляется огромное количество новых данных (в среднем 30000 новых постов и 480000 комментариев к ним)?

О. Кольцова и С. Кольцов в работе [2] показали успешность применения алгоритма LDA (латентного размещения Дирихле) для выявления тематической структуры больших совокупностей текстов Живого Журнала, что позволяет решить вторую проблему: автоматическим путем сформировать темы в исследуемой совокупности текстов и отобрать только те тексты, которые привязаны к

интересующим темам исследования (т. е. таким образом существенно сократив количество текстов для ручного анализа). Решению второй проблемы посвящено данное исследование.

1 Цели и задачи

Цель данной работы состоит в том, чтобы сравнить тематику и другие характеристики «популярных» и «обычных» блогеров; под первыми здесь понимаются блогеры, занимающие верхние позиции в рейтингах популярности.
Перед началом исследования мы сформировали две гипотезы:

1. топовые блогеры, ориентированные на публичность и лидерство в формировании общественного мнения, больше пишут для широкой аудитории и на темы, представляющие общественный интерес, тогда как «обычные» блогеры больше пишут о приватных и рекреативных вопросах для своих личных знакомых;
2. Обычные блогеры, не будучи профессионалами, в отличие от популярных блогеров, характеризуются меньшей активностью и смещением этой активности на выходные дни, тогда как популярные блогеры пишут, в основном, по будням.

2 Реализация

Данные собраны при помощи разработанного в Лаборатории интренет-исследований программного обеспечения BlogMiner, который позволяет закачивать и хранить посты и комментарии из Живого Журнала вместе с метаданными о аккаунте, времени и дате написания поста или комментария и ссылки на этой комментарий в Живом Журнале.
Выборка включила в себя все посты за месячный период (с 14 сентября по 14 октября 2013 года), созданные первыми 2000 блогерами по рейтингу «Социальный капитал» Живого Журнала[1] и 20000 случайных блогеров, представленных в данном рейтинге с 2001 по 150000 места; всего – 298967 постов и 2800154 комментариев. Предварительные исследования выявили, что количество постов резко падает после 150000 места, что обуславливает выбор данного ранга в качестве нижнего порога. Также ранее было показано, что 20000 нетоповых блогеров создают приблизительно столько же постов, как и первые 2000 блогеров, поэтому количество случайных блогеров было ограничено данным числом.
Автоматическое выделение тем, присутствующих в коллекции постов топовых и нетоповых блогеров, проводилось с помощью алгоритма латентного размещения Дирихле с сэмплированием Гиббса [4] с помощью разработанного в лаборатории программного обеспечения TopicMiner (http://linis.hse.ru/soft-linis). Необходимым параметром для алгоритма является задаваемое вручную количество

[1] http://www.livejournal.com/ratings/users/authority/?country=cyr

тем. После ряда тестов на основе непараметрического метода скачков [5] было определено, что оптимальным для нашей выборки является значения данного параметра равное 120 темам.
В результате тематического моделирования на основе данного алгоритма мы получили две матрицы: матрица, содержащая распределения слов по темам и матрица распределений документов по темам, при этом каждый столбец матрицы означает отдельную тему. Элементы матриц в каждой теме были отсортированы по убыванию. Таким образом были выделены 100 наиболее вероятностных документов по всем темам, которые были переданы двум кодировщикам для присвоения им ярлыков (интерпретации содержания тем)

3 Результаты

Проанализировав две группы пользователей Живого Журнала с точки зрения тематической структуры и других социологических показателей мы получили:

- Данные нетоповых блогеров сильно зашумлены: 25% от всех постов (42300) в случайной выборке были написаны одним аккаунтом спамерского происхождения. Данный феномен удалось обнаружить при помощи построения графика распределения количества постов на пользователя, а также распределения количества постов по дням недели (все эти тексты были выложены в Живой Журнал в период с 9 по 14 октября 2013 года).
- Активность нетоповых блогеров гораздо ниже, чем у топовых: большинство нетоповых блогеров, которые вообще имеют посты за исследуемый период, имеют по одному посту, в то время как у топовых блогеров этот показатель равен 40-60 на аккаунт. Кроме того, почти три четверти постов «обычных» пользователей не получили ни одного комментария, у топовых блогеров не получили комментариев менее трети постов. Кроме того, в постах топовых блогеров нередко встречаются дискуссии не менее чем из 10 комментариев, у нетоповых блогеров таких дискуссий крайне мало.
- Обычные блогеры склонны больше писать в будние дни, чем в выходные, причем примерно в той же мере, в которой и популярные блогеры; таким образом, вторая часть гипотезы 2 не подтвердилась.
- Для сравнения тематического состава топовых и нетоповых блогеров нами были просуммированы вероятности отнесения постов топовых и нетоповых блогеров к тем или иным темам. Затем была выделена доля каждой темы в общем весе тем у топовых и нетоповых блогеров по отдельности и было установлено, что в обеих выборках распределение тем практически идентично. Таким образом, мы не можем подтвердить нашу гипотезу о том, что топовые блогеры больше пишут на социально-политические темы, а нетоповых блогеров больше волнуют темы отдыха и личных взаимоотношений. При проведении тематического моделирования из изучаемой выборки не были удалены спамерские аккаунты, и наибольшее различие в тематике обеспечивается именно ими.

- В целом, тематическая структура постов топовых и нетоповых блогеров сходна с результатами предыдущих исследований авторов [5], и отличается в основном событийными темами.

4 Заключение

В результате проведенного нами исследования мы можем утверждать, что «обычные» и «популярные» блогеры, в равной степени интересуются как социально-политическими вопросами, так и личной и рекреационной сферами. В отсутствии различия в тематической структуре можно было бы советовать социологам использовать тексты не только популярных блогеров, но и обычных пользователей Живого Журнала для выявлении общественного мнения в блогосфере. Однако, меньшая активность и б*о*льшая зашумленность данных не позволяет этого сделать. Можно также сделать вывод о целесообразности расширения совокупности текстов для изучения онлайнового общественного мнения путем присоединения к общей выборки комментариев к постам популярных блогеров (которые в основном создают нетоповые блогеры).

Благодарности. В данной научной работе использованы результаты проекта «Социально-политические процессы в Интернете», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2013 году.

Список источников

1. González-Bailón S., Banchs R.E., Kaltenbrunner A. Emotions, Public Opinion, and U.S. Presidential Approval Rates: A 5-Year Analysis of Online Political Discussions // Human Communication Research. - Newark, DE, 2012. Vol. 38. № 2. P. 121–143
2. Koltsova O., Koltcov S. Mapping the public agenda with topic modeling: The case of the Russian livejournal // Policy & Internet. – UK: Wiley-Blackwell, 2013. Vol. 5. № 2. P. 207–227
3. Etling B. et al. Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization. - Rochester, NY: Social Science Research Network, 2010.
4. Griffiths T.L., Steyvers M. (2004) Finding scientific topics // Proceedings of the National Academy of Sciences, 101. P. 5228–5235.
5. Sugar, C.A. and James, G. M. Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach, Journal of the American Statistical Association, 2003; 98(463): 750-763.

Vox Populi Online: The Comparison of Posts' Structure and Topics Among the "Regular" and "Popular" Bloggers on LiveJournal

Svetlana Alekseeva, Olesya Koltsova, Sergei Koltsov

Higher School of Economics, Saint Petersburg, Russia
{salexeeva,ekoltsova,skoltsov}@hse.ru

Abstract. The paper is devoted to comparison of topical structure and basic statistical parameters among the "regular" and "popular" bloggers on LiveJournal. The study has shown a significant topical similarity between both of the user groups. The hypothesis that "popular" bloggers are more interested in social and political topics rather than "regular" ones has been rejected. The discovered difference between the groups is in "regular" users' lesser activity and increased data noise among them.

Keywords: topic modeling, LDA, LiveJournal, public opinion.

Оценка параметров хаотического процесса с помощью UKF-фильтра для построения прогноза

Елена Малютина, Владимир Иванович Ширяев

Южно-Уральский государственный университет, Челябинск, Россия
e.mankevich@gmail.com, vis@prima.susus.ac.ru

Аннотация. Решается задача идентификации хаотической компоненты временного процесса в условиях малого числа доступных наблюдений и когда реализация процесса единственна. Полученная в результате решения поставленной задачи аппроксимация используется для построения прогноза исследуемого процесса. В качестве модели хаотического сигнала предлагается использовать разложение по системе хаотических процессов, описываемых логистическим отображением. При этом параметры логистического отображения (состояние системы на каждом предыдущем шаге, лямбда) известны неточно и оцениваются с помощью UKF-фильтра. Поступление новых наблюдений позволяет найти новые оценки параметров и скорректировать модель, на основании которой строится дальнейший прогноз с сохранением заданной точности.

Ключевые слова: детерминированный хаос, нелинейная динамика, прогнозирование, короткий временной ряд, хаотическое моделирование

1 Введение

Задача идентификации хаотического сигнала [1, 2, 5, 6] имеет множество приложений в технических, информационных и социально-экономических системах, например, восстановление модели внешний возмущений на управляемый объект, построение модели сигнала на выходе высокочувствительного датчика [5, 6], повышение точности краткосрочных прогнозов [5, 6, 11, 12]. Актуальность исследования заключается в том, что при решении задач обработки сигналов, в которых содержится хаотическая составляющая [1, 14], сложность состоит в том, что выборка для обработки имеет малую длину, реализация процесса единственна и нет информации о вероятностных распределениях ошибок. Применение линейных моделей в случаях, когда шумы имеют фрактальную природу, не обеспечивают приемлемую точность [6]. Много исследований хаотических процессов посвящено методам реконструкции динамических систем по экспериментальным данным, основанным на применении нейронных сетей. Но для обучения нейронных сетей требуется большой объем выборки, что не всегда можно получить в реальных условиях. В связи с этим становится акту-

альной разработка соответствующих алгоритмов фильтрации для хаотических процессов [6,7].

В связи с тем, что логистическое отображение, которое используется для аппроксимации исследуемого процесса, является нелинейной функцией, параметры логистического отображения предлагается оценивать с помощью ансцентного фильтра Калмана (UKF) [13], предложенного оксфордскими учеными Джулье С. и Ульманом Д. в 1996 году.

2 Постановка задачи

Для построения модели хаотического процесса y_k, $k=\overline{1,N}$ предлагается использовать разложени по системе процессов, заданных нелинейными отображениями

$$x_{k+1}^{(i)} = f_i\left(x_k^{(i)}, \lambda_i\right),\ k=0,1,\ldots,N-1,\ i=1,2,\ldots,n, \tag{1}$$

где функции f_i, $i=1,2,\ldots,n$ определены на единичном интервале, то есть $f_i:[0,1]\to[0,1]$. В качестве примера таких процессов рассматриваются логистические отображения [3, 8, 9]

$$x_{k+1}^{(i)} = \lambda_i x_k^{(i)}\left(1-x_k^{(i)}\right),\ k=0,1,\ldots,N,\ i=1,2,\ldots,n, \tag{2}$$

хаотические решения которых возникают при $x_0^{(i)}\in(0,1),\ \lambda_i\in(\lambda_\infty,4]$, где $\lambda_\infty\approx 3.57$.

Для проверки работы ансцентного фильтра Калмана рассматривается модельный пример [4] при числе базисных процессов $n=1$ и отсутствии шума в системе, параметр $a=1$:

$$\begin{aligned} y_k &= ax_k+\eta_k,\\ x_k &= \lambda x_{k-1}\left(1-x_{k-1}\right),\ k=1,2,...,N, \end{aligned} \tag{3}$$

где $x_k\in R$ – хаотическое решение системы, $y_k\in R$ – вектор измерений, λ – параметр логистического отображения, $\eta_k\sim N(0,\sigma)$ – шумы в измерениях.

3 Результаты оценивания

Для оценки процесса x_k, заданного системой (4), на каждом шаге наблюдаемого процесса y_k, $k=\overline{1,N}$ и параметра логистического отображения λ был использован ансцентный фильтр Калмана [13]. Значения параметров логистического отображения $x_0=0.3,\ \lambda=3.69$ заданы с ошибкой 10% и составляют

$x_0 = 0.33,\ \lambda = 4.06$. Отношение сигнал/шум (SNR) наблюдаемого процесса y_k составляет 10 дБ (рис.3). Соответственно СКО σ_x для шума $\eta \sim N(0, \sigma_\eta)$ находится из соотношения $SNR = 20\log_{10}\frac{\sigma_x}{\sigma_\eta}$. В результате получены следующие оценки x_k (рис.1).

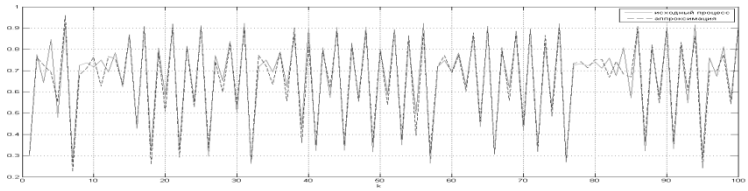


Рис. 1. Исходный процесс (–) и его аппроксимация (- -)

При количестве измерений N=60 была получена оценка параметра $\lambda = 3,65$ с абсолютной ошибкой, не превышающей 0,04 (рис.2).

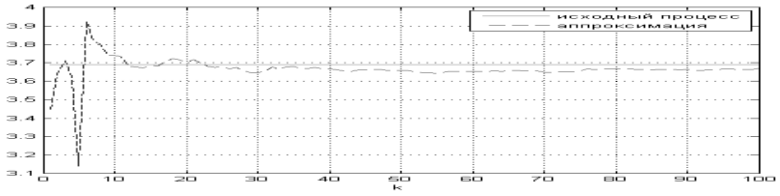


Рис. 2. Оценка параметра λ (- -) и его истинное значение

Полученные на шаге $k = 60$ оценки x_k и λ используем для построения прогноза на $L = 9$ шагов вперед (рис.3). Относительная ошибка прогноза на 4 шага вперед не превысила 2%, на 8 шагов – 10%. Ошибка прогноза на 9 шагов резко возросла до 20%.

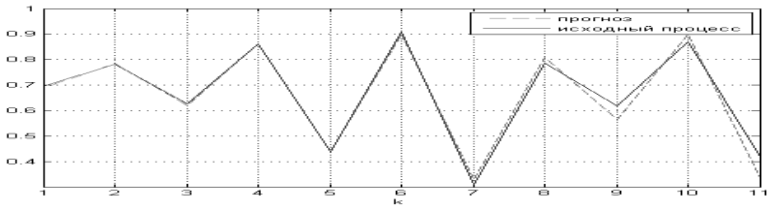


Рис. 3. Прогноз исходного процесса

4 Заключение

Приведен подход к построению модели хаотического процесса по малому числу (N=60) измерений при единичной реализации. Для оценки параметров был использован ансцентный фильтр Калмана (UKF). С помощью предложенного метода были получены оценки параметров логистического отображения при SNR=10 дБ. Сходимость оценки параметра λ была получена на коротком отрезке выборки N=60. При этом абсолютная ошибка оценивания параметра λ

не превысила 0,04. Относительная ошибка прогноза до 8 шага не превысила 10%.

Список источников

1. Андреев Ю.В., Дмитриев А.С., Ефремова Е.В. Разделение хаотических сигналов при наличии шума // Радиотехника и электроника. – 2001. – Т. 46, № 12. – С. 1460–1470.
2. Гришин И.В., Манкевич Е.И., Телегина К.В., Шелудько А.С., Ширяев В.И. О решении задач параметрической идентификации процессов с хаотической динамикой // Вестник ЮУрГУ. Серия компьютерные технологии, управление, радиоэлектроника. – 2008. – № 3. – С. 44–50.
3. Малютина Е.И., Соколова Т.Е., Ширяев В.И. Об одном подходе к аппроксимации и прогнозированию хаотических процессов // Актуальные проблемы автоматизации и управления. – Челябинск: Изд. центр ЮУрГУ, 2013. – С. 81–85.
4. Малютина Е.И., Ширяев В.И. Анализ коротких временных рядов на основе теории детерминированного хаоса // XXVII Международная научно-практическая конференция “Экономико-правовые и управленческие методики преодоления социальных кризисов”. 28 июня – 6 июля 2012 г. Международная академия наук и высшего образования (МАНВО; Лондон, Великобритания).– 2012– С. 163-167.
5. Манько Н.Г., Шалимов Л.Н., Шестаков Г.В., Штыков А.Н., Шелудько А.С., Ширяев В.И. Повышение точности оценок в алгоритме обработки измерений на выходе волоконно-оптического гироскопа с помощью применения моделей детерминированного хаоса // Актуальные проблемы автоматизации и управления. Тр. науч.-практ. конф. - Челябинск: ЮУрГУ, 2013. - С.43-46.
6. Разработка алгоритмов обработки измерительной информации и анализ точности волоконно-оптического гироскопа ВОГК-2 и его модификаций: отчет о НИР (инж. записка): ОКБ/103-11 2011058 / ЮУрГУ; рук. В.И. Ширяев; исполн.: А.С. Шелудько [и др.]. – Челябинск, 2011. – 110 с.
7. Тратас, Ю. Г. Оптимальный фильтр для неизвестного сигнала / Ю. Г. Тратас // Успехи современной радиоэлектроники. – 2013. – № 3. – С. 84-89.
8. Тюкин И.Ю., Терехов В.А. Адаптация в нелинейных динамических системах. – М.: ЛКИ, 2008. – 384 с.
9. Шелудько А.С., Ширяев В.И. Об одном подходе к построению модели хаотической компоненты временных процессов на коротких интервалах времени // Экстремальная робототехника. – СПб.: Политехника-сервис, 2010. – С. 101–108.
10. Шелудько А.С., Ширяев В.И. Совместное использование фильтра Калмана и минимаксного фильтра в задаче оценивания параметров модели хаотического процесса // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2012. – № 35. – С. 59–64.
11. Ширяев В.И. Финансовые рынки. Нейронные сети, хаос и нелинейная динамика. М.: Либроком, 2009. 232 с.
12. Яковлев В.Л., Яковлева Г.Л. Лисицкий Л.А. Модели детерминированного хаоса в задаче прогнозирования тенденций финансовых рынков и их нейросетевая реализация // Информационные технологии. 2000. № 2. С.46–52.
13. Julier S.J., Uhlmann J.K. A new extension of the Kalman Filter to nonlinear systems // In Proc. of AeroSense: The 11th Int. Symp. on Aerospace/Defence Sensing, Simulation and Controls. – Orlando, FL, USA, 1997. – P. 132-193.
14. Skiadas C.H. Chaotic modeling and simulation. – Boca Raton: CRC Press, 2008. – 364 p.

Parameter Estimation of Chaotic Process Using UKF and Time Series Forecasting

Elena Malyutina, Vladimir I. Shiryaev

South Ural State University, Chelyabinsk, Russia
e.mankevich@gmail.com, vis@prima.susus.ac.ru

Abstract. The study is devoted to the identification of the chaotic component of the time process in terms of small number of available observations and one process implementation. The approximation obtained from the solution of the problem is used to predict the investigated process. Decomposition in the system of chaotic processes described by the logistic map is used as a model of chaotic signal. Moreover, the parameter of the logistic map and the state of the system of each previous step are known inaccurately and are estimated using the unscented Kalman filter (UKF). Supply of new observations allows finding new parameter estimations and adjusts the model based on which to build further forecast maintaining the specified accuracy.

Keywords: deterministic chaos, non-linear dynamics, forecasting, short time series, chaotic modeling.

Использование семантического анализа текстов для поиска специалистов

Игорь Захлебин

НИУ ВШЭ, Москва, Россия
zahl.igor@gmail.com

Аннотация В работе предложен метод семантического поиска специалистов по набору составленных ими текстов. Описан формат запросов, позволяющий определять набор искомых компетенций. Разработаны алгоритмы построения и сравнения семантических представлений фрагментов текстов на естественном языке. На основе предложенной модели разработан и испытан прототип поисковой системы ExpSearch-1 (Experts Search, версия 1).

Ключевые слова: поиск специалистов, семантический анализ, теория К-представлений, естественно-языковые запросы.

1 Введение

Перед современными компаниями остро стоит проблема поиска квалифицированных специалистов. При этом поиск приходится осуществлять не только среди кандидатов на открывшиеся позиции, но и среди собственных сотрудников, например, для устранения нештатных ситуаций [1]. Поэтому для повышения эффективности бизнеса повсеместно разрабатываются автоматические системы, позволяющие ускорить и качественно улучшить этот процесс. Так, компания IBM менее чем за 6 лет сэкономила около $500 миллионов благодаря внедрению собственной системы поиска персонала [2].

Среди методов, применяющихся для поиска специалистов, наиболее популярным остается поиск по ключевым словам. Как правило, менеджер по персоналу, имеющий базу резюме специалистов, с помощью специального программного обеспечения осуществляет поиск по названию профессии, названиям технических средств и/или профилю образования специалиста. При этом, даже если запрос составлен удовлетворительно:

- при поиске не будут учтены смысловые отношения между словами;
- поисковая выдача будет различаться для запросов с одинаковым значением, но составленных по-разному (даже при учете синонимии понятий);
- оказывается невозможным алгоритмически определить, является ли конкретный пункт выдачи действительно релевантным запросу.

Для устранения перечисленных недостатков в данной работе предлагается модель системы поиска, использующая семантический анализ текстов и оценку релевантности результатов, основанную на сравнении семантических представлений фрагментов текстов.

2 Структура разработанной системы

Для реализации нового подхода к семантическому поиску специалистов была разработана система-прототип ExpSearch-1 (Experts Search, версия 1). Ее структура изображена на рис. 1.

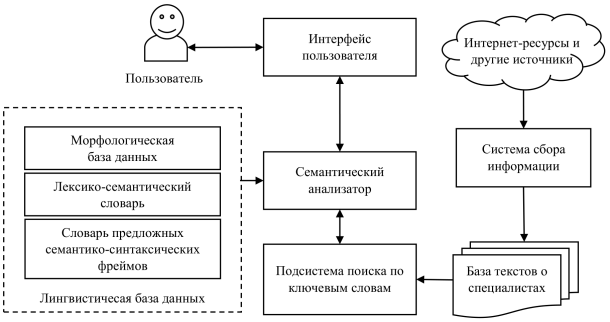


Рис. 1. Схема разработанной системы семантического поиска

В систему загружаются полные тексты с информацией о специалистах (анкеты, резюме, профессиональная переписка и т.п.), которые группируются по принадлежности к соответствующим специалистам.

Для поиска пользователь вводит запрос в виде набора словосочетаний, каждое из которых определяет одну искомую компетенцию [2]. Запрос состоит из существительных с возможным использованием предлогов, прилагательных и числительных. Запрос такого вида позволяет задать, например:

- область знания (эпизодическая логика, управление рисками);
- модель, теорию, понятие (модель Эрроу-Дебре, дефлятор ВВП);
- инструментальное средство (среда SPSS, пакет MatLab);
- умение или навык (обработка древесины, разработка под iOS).

Система ищет специалистов, у которых в связанных с ними текстах присутствуют релевантные словосочетания. Чем большему числу критериев удовлетворяет специалист, тем выше он располагается в ранжировании, выдаваемом системой.

3 Алгоритм построения семантических представлений

В основе работы системы лежит модифицированный алгоритм построения семантических представлений (далее – СП) и модель лингвистической базы данных, предложенные в книгах [3,4].

Построение СП фрагмента текста начинается с определения морфологических свойств его слов и приведения их к начальной форме. Затем к существительным применяется лексико-семантический словарь. По начальной

форме слова он сопоставляет ему семантическое значение sem (для равнозначных с точки зрения системы слов оно совпадает) и набор характеристик $st_1, \ldots, st_k$. Словарь содержит записи вида $(lec, sem, st_1, \ldots, st_k)$, где lec – базовая форма слова; sem – строка, обозначающая семантическое значение лексемы lec; $st_1, \ldots, st_k$ – различные семантические характеристики сущности, связанные с понятием sem; k – наибольшее возможное их число.

Далее к существительным применяется словарь предложных семантико-синтаксических фреймов, задающий связи между семантическими единицами, выделенными на предыдущем этапе. Он содержит записи вида $(prep, st_1, st_2, grc, rel)$, где $prep$ – необходимый предлог (может быть пустым); st_1, st_2 – семантические характеристики, которые можно связать с первым и вторым существительным в лингвистически правильном словосочетании «сущ.1 + $prep$ + сущ.2» соответственно; grc (grammatic case) – обозначение падежа второго существительного; rel – обозначение смыслового отношения. Существительные попарно проверяются на соответствие следующим условиям: первому существительному сопоставлен сорт sr_1, второму – sr_2, зависимое существительное находится в падеже grc, и между ними есть предлог $prep$. При удовлетворении всех условий для одной записи словаря считается, что между существительными установлено смысловое отношение rel из этой записи, дальнейшая сверка по словарю для этой пары прекращается.

Заметим, что на предыдущих шагах обрабатывались только существительные. Если имеется прилагательное или слово, ведущее себя как прилагательное, рассматриваются существительные, между которыми оно расположено. При совпадении с одним из них по роду, числу и падежу оно обозначается зависимым от него. При совпадении с обоими существительными прилагательное считается зависимым от последнего из них по порядку. В данных случаях устанавливается отношение rel – «свойство», а значение sem зависимой единицы – как начальная форма зависимого слова.

В результате выполнения алгоритма получается СП фрагмента текста – ориентированное дерево, в вершинах которого находятся семантические единицы sem, а ребра заданы отношениями rel.

4 Алгоритм поиска

Задачей алгоритма поиска является нахождение фрагментов в текстах о специалистах, имеющих СП (семантические представления), схожие с СП поискового запроса. Поэтому при поиске сначала строится СП запроса, а затем для каждого введенного пользователем словосочетания система составляет набор слов, парные вхождения которых в текст могут потенциально содержать между собой отношения как в СП запроса. Для этого в список включаются все слова из лексико-семантического словаря, которым могут быть сопоставлены единицы sem из СП запроса. Например, для словосочетания «маркетинг сбыта» может быть составлен следующий набор ключевых слов: «маркетинг», «маркетолог», «анализ рынка», «исследование рынка», «сбыт», «продажа».

По текстам, содержащимся в базе знаний, производится поиск по составленному набору ключевых слов. При нахождении хотя бы одного слова строится СП фрагмента текста вокруг него (границы определяются по таким символам, как точка, точка с запятой, табуляция, перенос строки и т.п.). Полученные представления группируются по специалистам, к текстам которых они относятся. Такой подход позволяет сохранить общую вычислительную сложность алгоритма низкой, так как процедура построения СП запускается только на предположительно релевантных фрагментах текстов.

Каждое СП можно упрощенно представить в виде набора триплетов вида (sem_1, rel, sem_2), то есть пар связанных семантических значений. Пусть $\mathbf{A}$ – набор триплетов, представляющих поисковый запрос, а $\mathbf{B}$ – аналогичный набор, представляющий СП, выделенные в текстах, связанных с одним специалистом. Тогда мерой релевантности специалиста будет величина $score = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A}|}$, находящаяся в отрезке $[0, 1]$. Для получения результирующего ранжирования специалисты упорядочиваются по убыванию показателя $score$, и их список возвращается пользователю как результат поиска.

5 Заключение

На основе предложенной модели на языке программирования Python был разработан прототип поисковой системы ExpSearch-1. В качестве тестовых данных в систему была загружена текстовая информация о более чем 7000 сотрудниках Высшей школы экономики (НИУ ВШЭ), взятая с официального сайта. В ходе испытаний система успешно выполнила поиск по набору тестовых запросов и дала по ним релевантные результаты.

В качестве направлений для продолжения работы предполагается усложнение формата поддерживаемых запросов и совершенствование алгоритма сравнения семантических представлений.

Список литературы

1. *Xiaodan Song и др.* ExpertiseNet: Relational and Evolutionary Expert Modeling. // SmallBlue Internet Edition (alpha) [Электронный ресурс]. URL: `http://smallblue.research.ibm.com/publications/ExpertiseNet_UM.pdf` (дата обращения: 10.03.2014).
2. *Arjen P. de Vries.* Expert Finding = Finding People + Assessing Expertise // Future Challenges in Expertise Retrieval, SIGIR 2008 Workshop, Singapore [Электронный ресурс]. URL: `https://app.box.com/s/9yqrk9zs6lc38gqpsi2j` (дата обращения: 10.03.2014).
3. *Фомичев В.А.* Формализация проектирования лингвистических процессоров – М.: МАКС Пресс, 2005.
4. *Fomichov V.A.* Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Series: IFSR International Series on Systems Science and Engineering, Vol. 27. Springer: New York, Dordrecht, Heidelberg, London, 2010.

Searching for Experts Using the Semantic Analysis of Texts

Igor Zahlebin

Higher School of Economics, Moscow, Russia
zahl.igor@gmail.com

Abstract. This paper presents a semantic method for searching for the experts. The method operates over a set of texts authored by themselves. The query format allowing one to define a set of the selected skills, and the algorithms for constructing and comparing the semantic representations are also presented. The ExpSearch-1 (Experts Search, version 1) system which is based on the present method has been developed and evaluated.

Keywords: experts' search, semantic analysis, K-representations, natural language queries.

Автоматическое порождение фраз естественного языка по OWL-модели, семантике и прагматике

Полина Сазонова

Новосибирский Государственный Университет, Новосибирск, Россия
psazonova@gmail.com

Аннотация Статья посвящена решению проблемы согласования пользовательской и программной онтологии. В ходе работы отработаны алгоритмы порождения фраз естественного языка по OWL-модели, семантике и прагматие. Данное решение можно использовать для повышения эффективности общения пользователей с виртуальными консультантами на сайтах, специализирующихся на продаже товаров и услуг.

Ключевые слова: онтология, OWL, virtual assistant, прагматика, семантика, NLP.

1 Введение

В настоящее время стремительно растет количество задач, связанных с обработкой данных, решаемых с помощью онтологий. Онтологии используются в теоретических исследованиях и практических разработках. По определению, формальная онтология предметной области представляет собой пару <S, σ>, где σ - это множество ключевых понятий, а S - множество аналитических предложений, описывающих смысл данных ключевых понятий [1]. Предложение называется аналитическим, если его значение истинности зависит только от смысла понятий, содержащихся в этом утверждении [2][3].

Онтологии применяются для построения интеллектуальных и экспертных систем. В настоящее время среди компаний, реализующих свои товары через Интеренет, стало популярным использование виртуальных помощников (virtual assistant). Их задачей является консультирование клиентов по вопросам, связанных с продаваемыми товарами, оказание помощи в выборе продукта, а в некоторых случаях и поддержание разговора на общие темы.

Виртуальный помощник ежедневно общается с разными людьми. Каждый потребитель (точнее, группа потребителей) имеют свое представление о предметной области и часто используют свою специфичную лексику. Поймут ли друг друга клиент и виртуальный помощник? Что следует делать, чтобы достигать наибольшего понимания? Здесь мы имеем проблему согласования программной и пользовательской онтологий [4]. В данной работе решается задача согласования этих двух онтологий для случая общения клиента и виртуального помощника.

Было установлено, что для решения обозначенной проблемы, необходимо найти "правила соответствия"между понятиями онтологии программной системы и онтологии пользовательских задач [4]. То есть виртуальному помощнику необходимо говорить на языке пользователя, а именно, осуществлять перевод информации с онтологии программной системы на онтологию конкретного пользователя (групп пользователей).

Сходным решением является технология, применяемая в чат-ботах (робот-собеседник), которая называется отзеркаливанием. Она заключается в перефразировании тех выражений, которые употребляет пользователь. Чат-бот отвечает на фразу собеседника подобной фразой, несколько измененной в соответствии с контекстом диалога. Таким образом, у пользователя создается ощущение, что его собеседник мыслит и ощущает так же, как он сам, что создает эмпатию. Однако, употребление только перефразированных выражений не может нести новую информацию. А в процессе диалога с виртуальным собеседником пользователю должна быть предоставлена новая информация, которая могла бы сподвигнуть его к совершению действий, например, покупке чего-либо. Причем новая информация должна быть представлена в том же стиле и при помощи такой же лексики, которую употребляет пользователь.

2 Разработка программной системы, осуществляющей порождение фразы естественного языка по модели, написанной на языке OWL

Решение проблемы согласования онтологий заключается в разработке такого виртуального помощника, который сможет использовать "правила соответствия"между понятиями онтологии программной системы и онтологии пользовательских задач.

Такая программная система порождает фразы естественного языка в полуавтоматическом режиме на основе семантики и прагматики. Понятие прагматики определяется с использованием подхода Фреге [5], где прагматика - это соответствие между синтаксисом и семантикой. В соответствии с этим подходом одна и та же семантика может выражаться при помощи разного синтаксиса. Вид используемого синтаксиса для заданной семантики зависит от пользователя, с которым происходит общение. Например, при общении с пожилыми людьми и с подростками семантически одинаковое выражение будет иметь различное синтаксическое воплощение.

В данной работе семантика для порождения фразы задается OWL-моделью.

Прагматика задается следующим образом. Мы делим пользователей на группы по половому признаку, возрастным категориям (молодежь, взрослый и пожилой) и по роду деятельности. Для каждой группы пользователей лексика задается при помощи онтологии, соответствующей именно этой группе, и словаря сленговых выражений, характерных для их профессиональной деятельности. При порождении фразы учитывается тип пользователя, тем самым создается контекст общения, комфортный для пользова-

теля. Это происходит за счет того, что лексика и способ ее употребления подбираются так, чтобы они были привычны пользователю. Проанализировав профиль человека, с которым происходит общение, можно получить информацию о его возрасте, деятельности и интересах, и далее сделать вывод о необходимости использования той или иной лексики и подобрать характеристики будущей фразы (её тональность и стиль). Программная система в процессе общения относит пользователя, с которым она ведет диалог, к определенной группе, и начинает использовать онтологии, соответствующие данной группе. Фраза строится на основе харктеристик тональности фразы и её стиля.

1. Тональность фразы задается наличием эмоционально окрашенной лексики. При порождении фразы, имеющей заданную тональность, подбираются соответствующие синонимы.
2. Стиль фразы определяется классом пользователей, к которому принадлежит собеседник. В соответствии с стилем выбираются нужная лексика, задаваемая соответствующей онтологией.

Построение фразы происходит на основе шаблонов с использованием принципов порождающей грамматики Хомского [6]. Каждое предложение состоит из именной группы и группы сказуемого. Именная группа состоит из определителя и определяемого, где определитель может опускаться. Группы могут содержать в себе прилагательные, наречия и т.д.

Согласование слов в предложении осуществляется за счет использования сторонних морфологических библиотек.

3 Заключение

В рамках данной работы были разработаны методы порождения фраз естественного языка по заданной OWL-модели, определяющей семантику фразы, и заданной прагматике фразы. Разработана программная система, реализующая данные методы.

Используя вышеописанный подход, можно решить проблему согласования пользовательской и программной онтологии и значительно повысить эффективность обмена знаниями между клиентом и виртуальным помощником. Виртуальный помощник при построении фразы учитывает не только модель ситуации, но и набор характеристик пользователя, с которым происходит общение, что гарантирует понимание сторон и увеличивает доверие к помощнику.

Список литературы

1. *Пальчунов, Д. Е.* Моделирование мышления и формализация рефлексии I: Теоретико-модельная формализация онтологии и рефлексии. // Философия науки. 2006. № 4 (31). С. 86-114.

2. *Carnap, R.* Meaning and Necessity. // A Study in Semantics and Modal Logic. 1956.
3. *Пальчунов, Д. Е.* Решение задачи поиска информации на основе онтологий. // Бизнес-информатика. 2008. № 1. С. 3-13.
4. *Пальчунов, Д. Е.*, *Целищев , В. В.* Проблема извлечения знаний в системе взаимодействия человека и компьютера (онтологии и пресуппозиции) // Философия науки. 2012. № 4. С. 20-35.
5. *Frege, G.* Über Sinn und Bedeutung / Zeitschrift für Philosophie und philosophische Kritik. s. 25–50. 1892.
6. *Chomsky, N.* Syntatic Structures — Berlin: Mouton de Gruyter, 2002, P. 119.
7. *Пальчунов, Д. Е.* Моделирование мышления и формализация рефлексии. Ч.2. Онтологии и формализация понятий. //Философия науки. 2008. № 2(37). С. 62–99.

Automatic Natural Language Generation Using an OWL Model, Semantics and Pragmatics

Polina Sazonova

Novosibirsk State University, Novosibirsk, Russia
psazonova@gmail.com

Abstract. This paper is focused on the problem of agreement between the ontology of a user and of a computer program. The natural language generation algorithms which use an OWL model, semantics and pragramtics have been studied. The present solution can be used to increase the interaction efficency between users and virtual assistants on websites specialized on selling goods and services.

Keywords: ontology, OWL, virtual assistant, pragratics, semantics, NLP.

Алгоритм множественного трекинга пешеходов

Роман Захаров

Самарский государственный аэрокосмический университет, Самара, Россия
roman.zakharovp@yandex.ru

Аннотация В данной работе приводится алгоритм множественного трекинга пешеходов. Трекинг основан на детекторе Part Based Detector, фильтре частиц и анализе траекторий движения пешеходов. Для разрешения коллизий с пересечением траекторий используется оптимизационный алгоритм решающий задачу о назначениях.

Ключевые слова: Part Based Detector (PBD), Histograms of Oriented Gradients (HOG), Viola-Jones, Kanade-Lucas-Tomasi, mean-shift, Particle Filters.

1 Введение

В настоящее время для решения многих практических задач используются системы компьютерного зрения (системы видеонаблюдения, системы помощи водителю и другие). В работе рассматривается задача нахождения локализации целевых объектов (пешеходов) и их сопровождение (трекинг).

Некоторые характеристики целевых объектов со временем изменяются, такие как, освещённость объекта, относительные размеры объекта при удалении или приближении объекта (пешехода) к камере. Объект может быть частично или полностью заслонён другими объектами. В данных условиях решение задачи локализации и сопровождения, становится задачей не тривиальной.

Для детектирования пешеходов широко используется метод гистограм ориентированных градиентов (Histograms of Oriented Gradients HOG[1]). Так же широко применяется метод Viola-Jones[2], который показывает хорошие результаты для детектирования человеческих лиц. Из алгоритмов трекинга довольно распространены трекинг на основе mean-shift и на основе оптического потока KLT (Kanade-Lucas-Tomasi). Так же в современных системах трекинг строится на основе фильтра Калмана и фильтра частиц (Particle Filters)[3].

В данной работе будет анализироваться система состоящая из детектора основанного на Part-Based-Detector (PBD[4]) и трекера основанного на фильтре частиц.

2 Описание задачи детектирования объектов

В настоящее время можно выделить два основных подхода к построению систем видеодетектирования: поиск и сопровождение областей движения; поиск и сопровождение уже обнаруженных объектов.

При реализации поиска и сопровождения областей движения обычно выделяют следующие этапы:

1. Получение кадра из видео последовательности
2. Выделение области движения на текущем кадре
3. Трекинг областей движения. Построение траектории движения
4. Разделение областей движения. В каждую область движения могут попасть несколько пешеходов. Для этого вычисляются параметры области движения, происходит поиск головы для разделения объектов внутри области движения.
5. Трекинг пешеходов. На данном этапе происходит построение и анализ траекторий движений пешеходов. Они могут пересекаться или идти довольно близко друг к другу из-за чего могут возникать коллизии с идентификацией объекта

В данный момент большинство систем работают по похожей схеме. Схема, основана на поиске и сопровождении областей движения, работает только с неподвижной камерой и как следствие с неизменным фоном. Данная схема - это типичная схема трекинга объектов различного типа.

В работе будет применяться несколько иная схема, которая позволит производить трекинг пешеходов при меняющемся фоне. Основные этапы следующие:

1. Получение кадра из видео последовательности
2. Применение детекторов к кадру видео последовательности. Для данного исследования был выбран детектор PBD обученный на базе данных PASCAL Visual Object Challenge-2007 (VOC2007). Наиболее популярным детектором для пешеходов является HOG
3. Анализ детектируемых областей. Отсечение областей по размеру
4. Трекинг пешеходов. Происходит построение и анализ траекторий движения пешеходов

Данная схема работает как при неподвижной камере так и на поворотных камерах. Что существенно расширяет область применимости данного метода. Это как стационарные камеры так и поворотные камеры, к примеру камеры прикреплённые к роботу или на автомобиль. Так же камеры на мобильных устройствах.

3 Описание алгоритма

Первым этапом алгоритма является применение натренированного детектора PBD к видео кадру. Детектор PBD[1] был выбран благодаря ста-

[1] Код детектора был взят с сайта `http://www.cs.berkeley.edu/~rbg/latent//`.

бильно высоким результатам детектирования объектов. На рис. 1 представлены детекты (прямоугольные области) которые выделил детектор после его выполнения.

Рис. 1. Результат применения детектора PBD к кадру видео последовательности

Далее идёт анализ детектируемых областей. Делаем простые отсечения по размеру детекта (прямоугольной области). Детекты размеры или соотношение размеров которых больше или меньше порогового значения отбрасываются. Пороговое значение для детектов подбирается экспериментально.

Следующий этап - это построение и анализ траекторий пешеходов. На рис. 2 представлены возможные траектории для пешеходов.

Для построения траектории движения пешехода используется фильтр частиц. Для разделения траекторий при их пересечении используется оптимизационный алгоритм решающий задачу назначений (Венгерский алгоритм). Который позволяет разделить две близкие траектории. Данный алгоритм трекинга устойчив к поворотам камеры, и как следствие устойчив к изменяющемуся фону. Данная устойчивость в трекере достигается за счёт того, что анализ происходит не области движения, а детектируемой области выделенной с помощью детектора PBD.

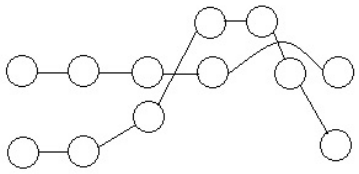


Рис. 2. Представлены возможные траектории пешеходов на последовательности кадрах

Список литературы

1. *Dalal, N. Triggs, W.* Histograms of Oriented Gradients for Human Detection. / IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05. — 2005. — Vol. 1(3). — P. 886–893.
2. *Viola P., Jones M.J.* Robust Real-Time Face Detection // International Journal of Computer Vision. 2004. №57(2). P. 137–154.
3. *Gustafsson F., Gunnarsson F., Bergman N. et al.* Particle Filters for Positioning, Navigation and Tracking // IEEE Transactions on Signal Processing. 2002. Vol. 2. Is. 2. P. 425–437.
4. *Felzenszwalb P. F., Girshick R. B., McAllester D., Ramanan D.* Object Detection with Discriminatively Trained Part Based Models // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010. V. 32. №9. P. 1627–1645.

Multi-Target Pedestrian Tracking Algorithm

Roman Zakharov

Samara State Aerospace University, Samara, Russia
roman.zakharovp@yandex.ru

Abstract. The general trend in the development of many video surveillance systems, traffic counting machines, security systems is the development of algorithms for tracking, which are very important in places with large numbers of people and vehicles, such as airports, city streets. In this paper we present an algorithm for multi-target tracking. Problem tracking is to build a trajectory of motion of targets on the input sequence of frames. Tracking is based on the detector Part Based Detector, the filter particles and analysis of the trajectories of pedestrians. To resolve conflicts with crossing trajectories used optimization algorithm solves the assignment problem.

Keywords: Part Based Detector (PBD), Histograms of Oriented Gradients (HOG), Viola-Jones, Kanade-Lucas-Tomasi, mean-shift, Particle Filters.

Об одном способе представления семантики текста

Ирина Николаевна Ефремова, Владислав Владиславович Ефремов

Юго-Западный государственный университет, Курск, Россия
Efremova-in@inbox.ru, v2@bk.ru

Аннотация. В работе рассматриваются вопросы «понимания» текстов на естественном языке. Предложена идея представления семантики текста с помощью аттракторов и дальнейшей работы со смыслом текста методами хаотической динамики.

Ключевые слова: семантика текста, понимание текста, аттрактор.

Список литературы

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
2. Хайдарова В. О некоторых видах аттракторов (на материале фразеоподсистемы языка Интернет-общения) // Проблемы истории, филологии, культуры. – 2008. – № 20. – С. 203–208.
3. Андреев Ю.В., Дмитриев А.С., Куминов Д.А. Хаотические процессоры// Успехи современной радиоэлектроники. - 1997. - №10. - С.4-26.

Toward a Method of Representing the Semantics of the Text

Irina N. Efremova, Vladislav V. Efremov

South-West State University, Kursk, Russia
Efremova-in@inbox.ru, v2@bk.ru

Abstract. This paper describes the problems of "understanding" of natural language texts. Authors proposed the idea of representing the semantics of the text using the attractors and further work with the semantics of text by methods of chaotic dynamics.

Keywords: semantics of text, text comprehension, attractor.

О представлении непрерывного оптического изображения в цифровом компьютере

Владислав Владиславович Ефремов, Ирина Николаевна Ефремов

Юго-Западный государственный университет, Курск, Россия
v2@bk.ru, Efremova-in@inbox.ru

Аннотация. В работе описаны разработанные способы для представления непрерывных изображений в цифровом компьютере, их основные функциональные возможности для обработки изображений.

Ключевые слова: оптический сигнал, дискретизация, интерполяция, квантование по уровню, группированная выборка.

Список литературы

1. Лемешко, Б.Ю. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход. — Новосибирск: Издательство НГТУ, 2011. – 888 с.
2. Лемешко, Б.Ю. Численное сравнение оценок максимального правдоподобия с одношаговыми и влияние точности оценивания на распределения статистик критериев согласия/ Б.Ю. Лемешко, Е.В. Чимитова // Заводская лаборатория. Диагностика материалов. - 2003. - Т.69. – С. 62-68.
3. Федосов, В.П. Формирование оптического изображения с помощью матричного фотоприёмника// Зарубежная радиоэлектроника. - 2001. - №9. - С. 59-63.
4. Цифровая обработка сигналов и изображений в радиофизических приложениях / Под ред. В.Ф. Кравченко. – М.: ФИЗМАТЛИТ, 2007. – 544 с.
5. Горшков, А.С. Цифровая обработка сигналов: атомарные функции и теория чисел. – М.: Машиностроение, 1994. – 224с.

Toward the Representation of Continuous Optical Images in a Digital Computer

Vladislav V. Efremov, Irina N. Efremova

South-West State University, Kursk, Russia
v2@bk.ru, Efremova-in@inbox.ru

Abstract. This paper describes the methods for representing continuous images in a digital computer developed by authors, their basic functional abilities for image processing.

Keywords: optical signal, discretisation, interpolation, quantization by level, grouped selection.