Vasyl Lytvyn
Natalia Sharonova
Thierry Hamon
Victoria Vysotska
Natalia Grabar
Agnieszka Kowalska-Styczen
(Eds.)

ISSN 2523-4013

COMPUTATIONAL LINGUISTICS AND INTELLIGENT SYSTEMS

Proceedings of the 2nd International Conference, COLINS 2018. Volume II: Workshop

Lviv, Ukraine
June, 2018

Lytvyn, V., Sharonova, N., Hamon, T., Vysotska, V., Grabar, N., Kowalska-Styczen, A., (Eds.): Computational Linguistics and Intelligent Systems. Proc. 2nd Int. Conf. COLINS 2018. Volume II: Workshop. Lviv, Ukraine, June 25-27, 2018, ISSN 2523-4013, colins.in.ua, online

This volume represents the proceedings of the Workshop Conference, with Posters and Demonstrations track, of the 2nd International Conference on Computational Linguistics and Intelligent Systems, held in Lviv, Ukraine, in June 2018. It comprises 19 contributed papers that were carefully peer-reviewed and selected from 36 submissions. The volume opens with the abstracts of the keynote talks. The rest of the collection is organized in two parts. Parts II contain the contributions to the Workshop COLINS Conference tracks, structured in two topical sections: (I) Computational Linguistics; (II) Intelligent Systems.

Preface

It is our pleasure to present you the proceedings of the Workshop Conference of COLINS 2018, the second edition of the International Conference on Computational Linguistics and Intelligent Systems, held in Lviv (Ukraine) on June 25-27, 2018.

The main purpose of the CoLInS conference is a discussion of the recent research results in all areas of Natural Language Processing and Intelligent Systems Development.

The conference is soliciting literature review, survey and research papers comments including, whilst not limited to, the following areas of interest:

- mathematical models of language;
- machine learning;
- discourse analysis;
- segmentation, tagging, and parsing;
- speech recognition;
- sentiment analysis and opinion mining;
- text categorization and topic modeling;
- text mining;
- information retrieval;
- artificial intelligence;
- information extraction;
- statistical language analysis;
- text summarization;
- data mining and data analysis;
- computer lexicography;
- social network analysis;
- question answering systems;
- web and social media;
- NLP applications;
- machine translation;
- intelligent text processing systems;
- memory systems and computer-aided translation tools;
- computer-aided language learning;
- corpus linguistics.

The language of COLINS Conference is English.

The conference took the form of oral presentation by invited keynote speakers plus presentations of peer-reviewed individual papers. There was also an exhibition area for poster and demo sessions. A Student section of the conference for students and PhD students run in parallel to the main conference.

This year Organizing Committee received 36 submissions, out of which 19 were accepted for presentation as a regular papers. The papers are submitted to the following tracks: discourse analysis (1 paper), data mining and data analysis (3 paper), web and social media (2 papers), segmentation, tagging, and parsing (1 papers), text categorization and topic modeling (3 paper), statistical language analysis (2 papers), text mining (1 paper), information retrieval (2 papers), artificial intelligence (3 papers), Natural Language Processing (3 papers), NLP applications (3 paper), corpus linguistics (2 papers), sentiment analysis and opinion mining (1 paper), computational lexicography (2 papers), automatic ontology building (1 paper), morphological analysis (1 paper), content analysis (3 papers), intelligent text processing systems (9 papers), intelligent computer systems building (9 papers) and problem of classification (1 papers). The papers directly deal with such languages: Ukrainian, Russian, French, English and Polish.

These papers and extended abstracts were published in this Volume II: Workshop of COLINS'2018 proceedings, Lviv, Ukraine, June 25-27, 2018, ISSN 2523-4013.

The conference would not have been possible without the support of many people. First of all, we would like to thank all the authors who submitted papers to COLINS 2018 and thus demonstrated their interest in the research problems within our scope. We are very grateful to the members of our Program Committee for providing timely and thorough reviews and, also, for being cooperative in doing additional review work. We would like to thank the Organizing Committee of the conference whose devotion and efficiency made this instance of COLINS a very interesting and effective scientific forum.

June, 2018

Vasyl Lytvyn
Natalia Sharonova
Thierry Hamon
Victoria Vysotska
Natalia Grabar
Agnieszka Kowalska-Styczen

Committees

General Chair

Yuriy Bobalo, Lviv Polytechnic National University, Ukraine

Steering Committee

Vasyl Lytvyn, Lviv Polytechnic National University, Ukraine
Natalia Sharonova, National Technical University "KhPI", Ukraine
Thierry Hamon, LIMSI-CNRS & Université Paris 13, France
Victoria Vysotska, Lviv Polytechnic National University, Ukraine
Natalia Grabar, CNRS UMR 8163 STL, France
Agnieszka Kowalska-Styczen, Silesian University of Technology, Poland

Program Chairs

Victoria Vysotska, Lviv Polytechnic National University, Ukraine
Dmytro Dosyn, Karpenko Physico-Mechanical Institute of the NAS of Ukraine
Olga Kanishcheva, National Technical University "KhPI", Ukraine

Proceedings Chair

Victoria Vysotska, Lviv Polytechnic National University, Ukraine
Vasyl Lytvyn, Lviv Polytechnic National University, Ukraine

Presentations Chair

Dmytro Dosyn, Karpenko Physico-Mechanical Institute of the NAS of Ukraine
Olga Lozynska, Lviv Polytechnic National University, Ukraine

Poster and Demo Chairs

Olga Kanishcheva, National Technical University "KhPI", Ukraine
Olga Lozynska, Lviv Polytechnic National University, Ukraine

PhD Symposium Chairs

Olena Levchenko, Lviv Polytechnic National University, Ukraine
Olga Lozynska, Lviv Polytechnic National University, Ukraine

IT Talks Chairs

Vasyl Lytvyn, Lviv Polytechnic National University, Ukraine
Olga Kanishcheva, National Technical University "KhPI", Ukraine

Local Organization Chairs

Vasyl Lytvyn (chair), Lviv Polytechnic National University, Ukraine
Victoria Vysotska (chair), Lviv Polytechnic National University, Ukraine
Olga Kanishcheva, National Technical University "KhPI", Ukraine

Publicity Chair

Vasyl Lytvyn, Lviv Polytechnic National University, Ukraine
Victoria Vysotska, Lviv Polytechnic National University, Ukraine

Web Chair

Victoria Vysotska, Lviv Polytechnic National University, Ukraine

Program Committees

MAIN Conference COLINS 2018

Yuriy Bobalo, Lviv Polytechnic National University, Ukraine)
Aleksandr Gozhyj, Petro Mohyla Black Sea National University, Ukraine
Anatoliy Sachenko, Ternopil National Economic University, Ukraine
Danuta Zakrzewska, Lodz University of Technology, Poland
Dmitry V. Lande, Institut for Information Recording of NAS of Ukraine
Dmytro Peleshko, IT Step University, Ukraine
Fadila Bentayeb, ERIC Laboratory, University of Lyon 2, France
Galia Angelova, Bulgarian Academy of Sciences, Bulgaria
Iryna Ivasenko, Karpenko Physico-Mechanical Institute of the NAS of Ukraine
Iryna Yevseyeva, Newcastle University, England
Klaus ten Hagen, University of Applied Science Zittau/Goerlitz, Germany
Lidia Pivovarova, University of Helsinki, Finland
Manik Sharma, DAV University, India
Michael Emmerich, Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
Michael Pokojovy, University of Memphis, TN, USA
Mikhail Godlevsky, National Technical University "KhPI", Ukraine
Natalia Grabar, CNRS UMR 8163 STL, France
Natalia Sharonova, National Technical University "KhPI", Ukraine
Nina Khairova, National Technical University "KhPI", Ukraine
Oksana Bihun, Mathematics University of Colorado, Colorado Springs USA
Oleg Bisikalo, Vinnytsia National Technical University, Ukraine
Oleg Garasym, Volvo IT, Poland
Olena Levchenko, Lviv Polytechnic National University, Ukraine
Olga Cherednichenko, National Technical University "KhPI", Ukraine
Scheller-Boltz Dennis, Vienna University of Economics and Business, Austria
Sergii Babichev, Jan Evangelista Purkinje University in Usti nad Labem, Chech Republic
Silakari Sanjay, Rajiv Gandhi Technical University, India
Svetla Boytcheva, Sofia University, Bulgarian Academy of Sciences, Bulgaria
Thierry Hamon, LIMSI-CNRS & Université Paris 13, France
Vasyl Lytvyn, Lviv Polytechnic National University, Ukraine
Victoria Bobicev, Technical University of Moldova, Moldova
Victoria Vysotska, Lviv Polytechnic National University, Ukraine
Viktor Mashkov, Jan Evangelista University in Ústí nad Labem, Czech Republic

Vitor Basto-Fernandes, University Institute of Lisbon, Portugal
Volodymyr Lytvynenko, Kherson National Technical University, Ukraine
Volodymyr Pasichnyk, Lviv Polytechnic National University, Ukraine
Waldemar Wojcik, Lublin University of Technology, Lublin, Poland
Wolfgang Kersten, Institut für Logistik und Unternehmensführung, Germany
Zoran Cekerevac, "Union – Nikola Tesla" University, Serbia

Posters and Demonstrations Track

Manik Sharma, DAV University, India
Olena Levchenko, Lviv Polytechnic National University, Ukraine
Dmytro Dosyn, Karpenko Physico-Mechanical Institute of the NAS of Ukraine
Oleksandr Gozhyi, Petro Mohyla Black Sea National University, Ukraine
Iryna Kalinina, Petro Mohyla Black Sea National University, Ukraine
Victor Gozhyj, Petro Mohyla Black Sea National University, Ukraine
Bohdan Rusyn, Humanities of Radom, Radom, Poland

Keynote Speakers

Natalia Sharonova, National Technical University "KhPI", Ukraine
Khrystyna Skopyk, Grammarly, Ukraine
Vasylyna Mytsak, N-iX, Ukraine
Thierry Hamon, LIMSI-CNRS & Université Paris 13, France
Natalia Grabar, CNRS UMR 8163 STL, France
Iryna Yevseyeva, Newcastle University, England
Vasyl Lytvyn, Lviv Polytechnic National University, Ukraine
Vitor Basto-Fernandes, University Institute of Lisbon, Portugal
Victoria Vysotska, Lviv Polytechnic National University, Ukraine
Dmytro Dosyn, Karpenko Physico-Mechanical Institute of the NAS of Ukraine
Michael Emmerich, Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
Olga Cherednichenko, National Technical University "KhPI", Ukraine

Additional Reviewers

Iryna Kalinina, Petro Mohyla Black Sea National University, Ukraine
Dmytro Dosyn, Karpenko Physico-Mechanical Institute of the NAS of Ukraine
Victor Gozhyj, Petro Mohyla Black Sea National University, Ukraine
Marianna Dilai, Lviv Polytechnic National University, Ukraine
Maxim Davydov, Lviv Polytechnic National University, Ukraine
Yevhen Burov, Lviv Polytechnic National University, Ukraine

Local Organization Committee

Vasyl Lytvyn, Lviv Polytechnic National University, Ukraine
Victoria Vysotska, Lviv Polytechnic National University, Ukraine
Olena Levchenko, Lviv Polytechnic National University, Ukraine
Olga Lozynska, Lviv Polytechnic National University, Ukraine
Dmytro Dosyn, Karpenko Physico-Mechanical Institute of the NAS of Ukraine

Sponsors

Lviv Polytechnic National University, Ukraine
http://www.lp.edu.ua/en

National Technical University "Kharkiv Polytechnic Institute"
http://www.kpi.kharkov.ua/eng/

Institut Galilée of Université Paris 13
https://galilee.univ-paris13.fr

Politechnika Śląska
https://www.polsl.pl/Strony/Witamy.aspx

Organizers

Lviv IT Cluster
https://itcluster.lviv.ua/en/

Kharkiv IT Cluster
http://it-kharkiv.com/en/

SoftServe
https://www.softserveinc.com/

Fortifier
www.4tifier.com

Envion Software
http://envionsoftware.com

PI-MINDS
http://pi-minds.com/en/

SSA Group
https://www.ssa.group

SYTOSS
https://sytoss.com

Conference Program COLINS-2018

Date	***25.06.2018***
Time	*Event*
12.00	***Registration*** (Room 124, Building 4)
13.00	***Conference Opening*** (Room 124, Building 4)
13.05	**Language modelling and its use cases** Khrystyna Skopyk (Grammarly)
13.30	**Modern information technologies in computer linguistics** Vasylyna Mytsak (N-iX)
13.55	**Uniform Method of Operative Content Management in Web Systems** Aleksandr Gozhyj, Lyubomyr Chyrun, Agnieszka Kowalska-Styczen, Olga Lozynska (Petro Mohyla Black Sea National University, Lviv Polytechnic National University, Silesian University of Technology)
14.20	**Web Content Support Method in Electronic Business Systems** Vitor Basto Fernandes, Victoria Vysotska, Michael Emmerich (University Institute of Lisbon, Lviv Polytechnic National University, Leiden University)
14.45	**The Typhlocomments Rules for Audiodescription System of the Video Content Formation for People with Visual Impairments** Andriy Demchuk, Olga Lozynska (Lviv Polytechnic National University)
14.10-15.00	***Coffee-break***
15.00	**Ontology Using for Decision Making in a Competitive Environment** Vasyl Lytvyn, Oksana Oborska, Victoria Vysotska, Dosyn Dmytro, Andriy Demchuk (Lviv Polytechnic National University, Karpenko Physico-Mechanical Institute of the NAS Ukraine, Silesian University of Technology)
15.25	**Methods of Information Resources Processing in Virtual Library** Bohdan Rusyn, Victoria Vysotska, Liubomyr Pohreliuk (Kazimierz Pulaski University of Technology and Humanities in Radom, Karpenko Physico-Mechanical Institute of the NAS Ukraine, Lviv Polytechnic National University)

15.50	**SEO Technology for Web Resource Processing** Lyubomyr Chyrun, Victoria Vysotska, Liliya Chyrun, Aleksandr Gozhyj, Iryna Kalinina (Lviv Polytechnic National University, Petro Mohyla Black Sea National University)
16.15	**Choosing the Method of Finding Similar Images in the Reverse Search System** Oleh Veres, Bohdan Rusyn, Anatoliy Sachenko, Ihor Rishnyak (Lviv Polytechnic National University, Kazimierz Pulaski University of Technology and Humanities in Radom, Karpenko Physico-Mechanical Institute of the NAS Ukraine, Ternopil National Economic University)
16.40	**Content Analysis of Text-based Information in E-commerce Systems** Vasyl Lytvyn, Victoria Vysotska, Lyubomyr Chyrun, Mariya Hrendus, Oleh Naum, (Lviv Polytechnic National University, Drohobych Ivan Franko State Pedagogical University, Silesian University of Technology)
17.05	**Content Formation Method in the Web Systems** Vasyl Lytvyn, Dmytro Dosyn, Michael Emmerich, Iryna Yevseyeva (Lviv Polytechnic National University, Karpenko Physico-Mechanical Institute of the NAS Ukraine, Leiden University, De Montfort University, Silesian University of Technology)
17.30-18.30	***Poster section and coffee-break***
Date	***26.06.2018***
Time	*Event*
10.00	***Conference Opening*** (Room 124, Building 4)
10.05	**Issues of Fact-based Information Analysis** Nataliia Sharonova, Anastasiia Doroshenko, Olga Cherednichenko (National Technical University "Kharkiv Polytechnic Institute")
10.45	**Ontologies Quality Assessment and Multicriteria Optimization** Vitor Basto Fernandes (University Institute of Lisbon)
11.10	**Portfolio Optimization in Non-financial Application Domains** Iryna Yevseyeva (De Montfort University)
11.35	**Multicriteria Analysis of Complex Networks** Michael Emmerich (Leiden University)

12.00	**Role of Machine Learning and Soft Computing Techniques in Medical Science** Manik Sharma (Computer Science and Applications DAV University)
12.25-13.00	***Poster section and coffee-break***
13.00	**Sentiment Analysis of the US and Ukrainian Presidential Speeches** Marianna Dilai, Yulia Onukevych, Iryna Dilay (Lviv Polytechnic National University)
13.25	**Studying Items Similarity for Dependable Buying on Electronic Marketplaces** Olga Cherednichenko, Maryna Vovk, Olga Kanishcheva, Mikhail Godlevskyi (National Technical University "Kharkiv Polytechnic Institute")
13.50	**The Quantitative Research of Scientific Texts at the Symbolic Level** Ihor Kulchytskyi, Ulyana Shandruk (Lviv Polytechnic National University)
14.15	**Recovery Gaps in Experimental Data** Roman Kaminskyi, Nataliia Kunanets, Volodymyr Pasichnyk, Antonii Rzheuskyi, Andrii Khudyi (Lviv Polytechnic National University)
14.40	**Automatic Detection of Temporal Information in Ukrainian General-language Texts** Natalia Grabar, Thierry Hamon (Université Paris 13)
15.05-15.30	***Poster section and coffee-break***

15.30-18.00	**Section I.** Computational Linguistics (Room 128, Building 4)	**Section II.** Intelligent Systems*)* (Room 124, Building 4)
15.30	**Detection of Gaps in Documentation Concerning Remote-piloted Aviation based on Content Analysis** Volodymyr Markiv, Mykola Mykyichuk, Oksana Markiv	**The Comparison of Holt – Winters and Box – Jenkins Methods for Software Failures Prediction** Vitaliy Yakovyna, Oleksandr Bachkai
15.55	**DFA Method for the Analysis of Long-Range Correlations: Application to Statistical Linguistics** L. B. Ivanitskyi, O. S. Kushnir	**Study Methods of Image Segmentation for Intelligent Surveillance Systems** Volodymyr Hrytsyk

16.20	**Distinguishing between Natural and Random Texts: a Statistical Measure Linked to Word Clustering** O. S. Kushnir, A. I. Kashuba, V. V. Yaremkiv	**Using Dynamic Neural Networks for Server Load Prediction** Pavlo Pukach, Volodymyr Hladun
16.45	**Embedding Speech Recognition Tools for Custom Software: Engines Overview** Arthur Dovbysh, Vladyslav Alieksieiev	**Peculiarities of Remote-Piloted Vehicles On-Board Navigation Complex Construction** Mykola Mykyichuk, Volodymyr Markiv
17.10	**Method for Automatic Collocation Extraction from Ukrainian Corpora** Maria Kuzmina, Svitlana Petrasova	**Reverse-search System of Similar or Identical Images** Oleh Veres, Yaroslav Kis, Vladyslav Kugivchak, Igor Rishniak
17.35	**The Information System for Identification of Content Set Based on Analysis of Similar Texts** Viktoriia Kovalchuk, Vasyl Lytvyn, Victoria Vysotska, Mariya Hrendus, Oleh Naum	**Extracting and Classification the Semi-Structured Data of Web-Systems** Irina Pelekh
18.00		**Influence of Readability on Popularization of Internet Resources** Taras Basyuk

18.25 ***Conference Closing***

Date	***27.06.2018***
Time	*Event*
10.00	***Coffee-break***
10.30-13.30	***Excuarsion***
15.10-15.30	***Coffee-break***
16.30-17.50	***Excuarsion***

18.00 ***Conference Closing***

Content

PART I.
KEYNOTE SPEAKERS TALKS

Ontology Using for Decision Making in a Competitive Environment

Vasyl Lytvyn[1], Oksana Oborska[2], Victoria Vysotska[3], Dmytro Dosyn[4], Andriy Demchuk[5]

[1-3,5]Lviv Polytechnic National University, Lviv, Ukraine
[1]Silesian University of Technologyty, Gliwice, Poland
[4]Karpenko Physico-Mechanical Institute of the NAS Ukraine

[1]Vasyl.V.Lytvyn@lpnu.ua,
[2]Oksana949@gmail.com,
[3]Victoria.A.Vysotska@lpnu.ua,
[4]Dmytro.Dosyn@gmail.com,
[5]Andrii.B.Demchuk@lpnu.ua

Abstract. A method of decision making system elaboration in competitive environment based on ontological approach was developed. The models based on a hinged Boyd-Moore automaton. State machine loops are the stages Boyd and filling processes, ontology editing and a search of relevant knowledge ontology. Possible transitions between the states of the machine and transfer parameters between them were determined. For a quantitative modeling of decision support process in a competitive environment, mathematical support and methods of domain-specific ontology in the Boyd cycle (OODA – observation, orientation, decision, action) were elaborated. Thus, for the military at an "Observations" stage intelligence domain ontology determining of the strengths and weaknesses of the enemy. At the stage of "orientation", ontological data are used for simulation of possible realizations of the battle and for optimal positioning of the forces. For "Resolution" stage, target distribution was developed using a method based on genetic algorithms, which helped to reduce the computational complexity of finding effective distribution targets. Ontology expertise based on descriptive logic is presented to increase the effectiveness of possible solutions.

Keywords: decision support system (DSS), ontology, knowledge database, Boyd cycle (OODA), observation, orientation, decision, action, genetic algorithms, expected value, probability.

1 Introduction

The effectiveness of modern armed forces largely depends on the professional level of command structure that, in turn, is defined by the degree of its automation [1]. Troop control automation can essentially enhance the combat capabilities and shorten the time supervisory units spend on operation planning and informing their command

subordinates. Automatic Control System (ACS) of Ukrainian Ground Forces tactical section is a totality of self-dependent bodies and command points that are equipped with computer-based decision support systems and means of communication that enable effective control of formations, units and subunits [2]. Decision Support System (DSS) is a central element of ACS. It enables military operations modeling, development of possible variants based on different criteria and transfer of recommendations to tactical section commanders. It is worth mentioning that a DSS is functioning in a competitive environment that involves several management entities that compete [3]. Modern approach to scientific modeling of decision support process in a competitive environment (like the military) one lies in the use of Boyd cycle that presupposes multiple recycling of four consecutive stages: observation, orientation, decision, and action [4]. This cycle is also abbreviated OODA. According to Boyd's hypothesis – the speed of the cycle and the accuracy of evaluation of its stages provide the advantage over the enemy and leads to the victory in warfare. When modeless warfare, we can distinguish between several important indexes that directly influence the result. These indexes of ground forces warfare modeling are: distance between troops, maneuver characteristics, practicability (motion resistance), target visibility (possibility of target detection), possibility of target destruction, sector of target search, density of fire means suballotmant on enemy's targets, number of shoots needed to destroy the target (spread characteristics, protection of the target, distance, etc.) [5-13].

In the majority of cases, the value of these indexes directly depends on operational and physical characteristics (OPC) of different types of armament and military equipment as well as the institutional and management structure of formations, units and subunits. That is why reliable software to save all this information that is required. Such information is to be stored in a knowledge database not in an ordinary database, because logical output is very important in the process of warfare modeling and can only be implemented on the base of knowledge of subject area. As far as the TTX OBT and organizational and staff structure is based on normative documents, the core for such knowledge database will be the ontology of Ukrainian Ground Forces [14]. The goal is to develop decision support methods in a competitive environment based on Boyd cycle by means of ontology use.

2 Materials and methods

Let us have a closer look at each stage of Boyd cycle in the process of interaction with domain-specific ontology and tasks that arise in this domain (Fig. 1).

Stage of observation gives the possibility to construct the ontology and to analyze it for relevant information that is needed in next stages of the OODA cycle. Ontology analysis is done on the basis of intelligence service data. Intelligence officers transfer information concerning enemy's assets that is found in ontology; information is processed and transferred to the tactical section commander.

The task of intelligence data processing is summarized as follows: intelligence officers provide messages (generally many-sided) concerning system X; it is required to define the system condition as precise as possible and distribute the possibility of

different conditions. Apart from that, while processing the intelligence data, only reliable messages should be analyzed, that means the conditions evaluation should be done in advance $x_1, x_2,..., x_k$ in $P_0(x_1), P_0(x_2), ..., P_0(x_k)$.

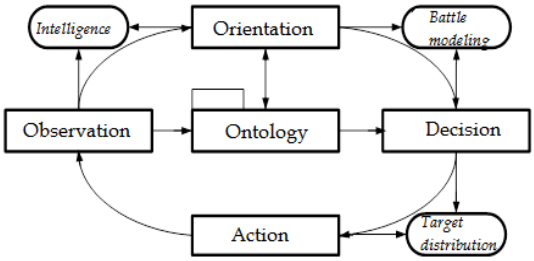


Fig. 1. Stages of Boyd cycle (OODA) and their interaction with ontology

Let us call these conditions preliminary ones to distinguish them from final ones received from intelligence officers. Apparently, final conditions depend on the set of messages collected by intelligence officers. Let us denote the set of messages by the letter $\tilde{S}$, and the final probability of conditions based on those messages as $P_p(x_1/\tilde{S}); P_p(x_2/\tilde{S}); ...; P_p(x_k/\tilde{S})$. These possibilities are conditional possibilities of conditions $x_1,...,x_k$, calculated on the event that the intelligence provided the set of messages $\tilde{S}$, that are calculated with the help of Bayes' formula

$$P_p(x_i/\tilde{S}) = \frac{P_0(x_i)P(\tilde{S}/x_i)}{P_0(x_1)P(\tilde{S}/x_1)+...+P_0(x_i)P(\tilde{S}/x_i)+...+P_0(x_k)P(\tilde{S}/x_k)}, \quad (1)$$

where $P(\tilde{S}/x_1)$ is a probability of the set of messages $\tilde{S}$ if a system is in state x_1; $P(\tilde{S}/x_2)$ is a probability of the set of messages if a system is in state x_2, etc. If no have any preliminary information on the system state are available, the conditions can be defined as equal:

$$P_0(x_1) = P_0(x_2) = ... = P_0(x_k) = \frac{1}{k}.$$

Let us assume that an intelligence officer delivers a message $\tilde{x}_1$. According to Bayes' formula (1), the probability of state x_1 equals to

$$P_p(x_1/\tilde{x}_1) = \frac{P_0(x_1)P(\tilde{x}_1/x_1)}{P_0(x_1)P(\tilde{x}_1/x_1)+P_0(x_2)P(\tilde{x}_1/x_2)}.$$

Obviously, $P_p(x_2/\tilde{x}_1) = 1 - P_p(\tilde{x}_1/x_1)$. Assume that an intelligence officer delivers a second message $\tilde{x}_2$. Then,

$$P_p(x_1/\tilde{x}_2) = \frac{P_0(x_1)P(\tilde{x}_2/x_1)}{P_0(x_1)P(\tilde{x}_2/x_1)+P_0(x_2)P(\tilde{x}_2/x_2)}.$$

If an intelligence officer does not deliver any message, than

$$P_p(x_1/\tilde{x}_0)=\frac{P_0(x_1)P(\tilde{x}_0/x_1)}{P_0(x_1)P(\tilde{x}_0/x_1)+P_0(x_2)P(\tilde{x}_0/x_2)}.$$

Intelligence is the most important element that guarantees advantage in warfare. Tactical intelligence is aimed at creation of favorable conditions to start a battle in an organized and well-timed manner and successful warfare conduction. That's why it is needed to develop software to transfer the information to the units quickly and effectively as well as to generalize the data concerning effective combat strength, location and enemy forces status, nature and intention of their action, strengths and weaknesses, level and type of equipment. To collect and process intelligence data, we have developed an Android application "Military Intelligence".

On the **stage of orientation** the strategy of action is defined. For this purpose, a modulus of simulation modeling of a battle was designed. It is described in details in section 4. This section describes the software the modulus functions on.

In the process of warfare modeling, we can distinguish between several parameters that influence the result. The parameters for warfare modeling for ground forces are the distance between troops, performance characteristics of mechanized infantry, the terrain: practicability (motion resistance), target visibility (possibility of target detection), possibility of target destruction, sector of target search, density of suballotmant of fire means on enemy's targets, number of shots needed to destroy the target (spread characteristics, protection of the target, distance, etc.) In the majority of cases, the value of these indexes directly depends on operational and physical characteristics (OPC) of different types of armament and military equipment as well as the institutional and management structure of formations, units and subunits. That is why reliable software is required to store and process all this information.

To determine which elements should be stored in ontology of knowledge database in DSS, let us analyze the mathematical models that are used in the process of warfare modeling. Mathematical warfare model is a two-set model $Q=\{q_1,q_2,...,q_n\}$ and $U=\{u_1,u_2,...,u_m\}$ that define the qualitative and quantitative structure of belligerent powers. For each element $q_i \in Q$, there exists a random multidimensional function $\zeta_i(t)=\zeta\left(\zeta_{i1}(t),\zeta_{i2}(t),...,\zeta_{ir(i)}(t)\right)$ for $T_0 \le t \le T_1$, where T_0 i T_1 respectively denote the start and the end periods of the battle, respectively. Random functions $\zeta_{i1}(t),\zeta_{i2}(t),...,\zeta_{ir(i)}(t)$ are referred to as parameters of the element q_i, l – implementation of a random function $\zeta_i(t)$: $\zeta_i^l(t)=\zeta_i\left(\zeta_{i1}^l(t),\zeta_{i2}^l(t),...,\zeta_{ir(i)}^l(t)\right)$. Random function section $\zeta_i(t)$ in a time set $T_0 \le t_z \le T_1$ is called the status of element q_i and is denoted as $C_i(t_z)$. The vector $\zeta_i^l(t_z)=\left(\zeta_{i1}^l(t_z),\zeta_{i2}^l(t_z),...,\zeta_{ir(i)}^l(t_z)\right)$ stands for the condition of element q_i in t_z for the 1st implementation and is written as $C_i^l(t_z)$. The collection $\{C_i^l(T_0)\}$ for all i=1,2,…,n describes the initial status of elements Q for the 1st implementation. Similarly, other elements are given by $U_j(j=1,2,...,m)$ and corresponding marks are entered:

$$\xi_j(t)=\xi_j\left(\xi_{j1}(t),\xi_{j2}(t),...,\xi_{jr(j)}(t)\right),\ \xi_j^l(t)=\xi_j\left(\xi_{j1}^l(t),\xi_{j2}^l(t),...,\xi_{jr(j)}^l(t)\right),$$

$$D_j(t_z)=\xi_j(t_z)=\left(\xi_{j1}(t_z),\xi_{j2}(t_z),...,\xi_{jr(j)}(t_z)\right),$$

$$D_j^l(t_z)=\xi_j^l(t_z)=\left(\xi_{j1}^l(t_z),\xi_{j2}^l(t_z),...,\xi_{jr(j)}^l(t_z)\right).$$

The collection $\{D_j^l(T_0)\}$ for all j=1,2,…,m is called the initial condition of a belligerent power U for the 1st implementation, and the collection $\{D_j^l(T_1)\}$ is an objective result of the battle for belligerent power U for the 1st implementation. $\{C_i^l(T_1)\}$ and $\{D_j^l(T_1)\}$ are called the together objective result of the battle for the 1st implementation, and $\{C_i^l(T_0)\}$ and $\{D_j^l(T_0)\}$ the initial condition of a battle for the 1st implementation. As parameters for chosen elements of the battle, the following random functions of a real argument t can be chosen: $\eta_1(t)$ – combat effectiveness; $\eta_2(t)$ – military position; $\eta_3(t)$ – speed; $\eta_4(t)$ – nature of action; $\eta_5(t)$ – ammunition amount. Detailed models of warfare that were used in the modulus of simulation modeling are described in the book "Mathematical models of warfare" edited by P. Tkachenko. To determine the importance of targets, the model of adaptive ontology designed by V. Lytvyn is used [14-18]. The importance of target is measured by the damage caused as a result of its destruction. The gradation of targets was determined after a survey of military sphere experts that were asked to assess the importance of ontology elements on a 10-point scale (1 – the important target – bullet pump, 10 – command post brigade) (1 – the importance of machine, gun target, 10 – the importance of team CP target). The importance of ontology element that sets the enemy target is calculated as the usage of experts evaluation, that is $W\in[1,10]$. Then, the enemy's most important target, as an ontology element, is calculated via the following formula:

$$C_{Z^*}=\arg\max_{C_Z}\left(\sum_{\tilde{C}_i\to C_Z}W_{\tilde{C}_i}+W_{C_Z}\right) \tag{2}$$

Decision-making is the third stage of the OODA cycle. If CP managed to shape only one real plan up to this stage, the decision is made whether to implement this plan or not. While improving the decision-making stage, we used target assignment problem. Target assignment is an operation which consists in assignments of a certain target to a certain fire weapon. It is necessary to find the optimal (best) target assignment by assigning to each cannon a certain target at which it should shoot (however, it is possible that one target will be attacked by several cannons). Assume here are n means of destruction at our disposal and we need to attack a dispersed group which consists of N targets. Each of the means of destruction can make only one shot and basically can shoot at each target, but with different effectiveness. Probability of hitting the j-th target by the i-th means of destruction equals to P_{ij}. In order to determine these probabilities, we use tables from regulations, which are given in the ontology of the Ukrainian Ground Forces. Probability data destruction of

enemy targets a certain way, taken from an ontology based on analysis of input parameters (appearance, permeability, speed, combat effectiveness, weather conditions). It is necessary to the find optimal (best) target assignment by assigning to each means of destruction a certain target at which it should aim (however, it is possible that one target will be attacked by several means of destruction). Such task is called an $n \times N$ target assignment. In order to solve a problem of target assignment, first of all, it is necessary to choose a performance indicator. Depending on shooting conditions, such indicator can be:

- the expectation of affected targets number (that doesn't team to be equal to the expectation);
- probability that each and every target will be affected, etc.

Indicator of target assignment effectiveness according to mathematical expectation is variable $M_n = M[X_n]$, where the random variable X_n is the number of targets hit. During shooting at the group target, the average number of hit targets equals to the sum of probabilities of certain basic targets (units) affection: $M_n = W_1 + W_2 + ... + W_N$, where W_1 – probability of the first target hitting; W_2 – probability of the second target hitting; W_N – probability of N-target hitting. Thus we obtain a task:

$$M_n = \sum_{j=1}^{N} W_j \to \max \tag{3}$$

Therefore, the target assignment with respect to the expectation, it is necessary to assign means of destruction to their targets in such a way that the sum of probabilities reaches a maximum. The simplest method of performing such an assignment is the brute force full search: all possible variants of cannon assignment to targets are searched and the one rendering the sum of the probabilities maximal is selected. Computational complexity of such algorithm is exponential and the number of possible combinations of target assignments equals A_N^n. If a number of possible variants is not very large, this search is possible to perform. However, in practice a number of targets and means of destruction is rather considerable. That is why in order to solve problem (3), it is proposed to use artificial intelligence methods, in particular genetic algorithms. In our case, equation (3) acts as a fitness function. If a number of means of destruction is much less than a number of targets $(n \prec N)$, targets are ranked in advance according to equation (2). For realization of the proposed approach, we chose relational database system implemented in MySQL, which contains information about our available means of destruction, proved enemy targets and probability matrix of destroying targets by a certain fire weapon. Chromosome is a vector, where the number of vector element is the key of means of destruction in database, and the meaning of element is the key of a target in the database. The optimal solution is than the chromosome with the largest value of fitness function. For modeling purposes, a certain number of chromosome generations were chosen. Experimental results showed that during generation of 30 breeds the best chromosome

close to the optimum target assignment was find. The target assignment module developed on the basis of genetic algorithms is a part of DSS.

The main advantage of the proposed approach is a considerable reduction of complexity of target assignment algorithm. Where the complexity of a complete enumeration is is exponential; the complexity of the genetic algorithm is linear with respect to the number of iteration. Thus, the process of decision-making is significantly accelerated. This is especially important when the events concerning target assignment take place in the real time. Although there is received solution of target assignment is not always optimal, it is close to the optimal one and the time advantage when computing the problem is considerable.

Action is the final stage of the cycle, which presupposes practical realization of the chosen idea or plan. Summing up everything previously mentioned, a model of Boyd cycle with the usage of ontology in the form of Moore automatic weapon (Fig. 2) was elaborated. Figure 2 displays: s_0 – initial state ("Observation" stage), s_1 – "Orientation" stage, s_2 – "Ontology editing", s_3 – "Search of relevant knowledge", s_4 – "Decision" stage, s_5 – "Action" stage; x_1 – data absent in ontology, x_2 – data about enemy, x_3 - ontological data (x_3' – for a decision, x_3'' – for an action), x_4 – environment assessment, x_5 – situation modeling, x_6 – data synthesis, x_7 – data analysis, x_8 – decision assessment, x_9 – data collection, x_{10} – proposed solution, x_{11} – ontology editing (new knowledge), x_{12} – environment.

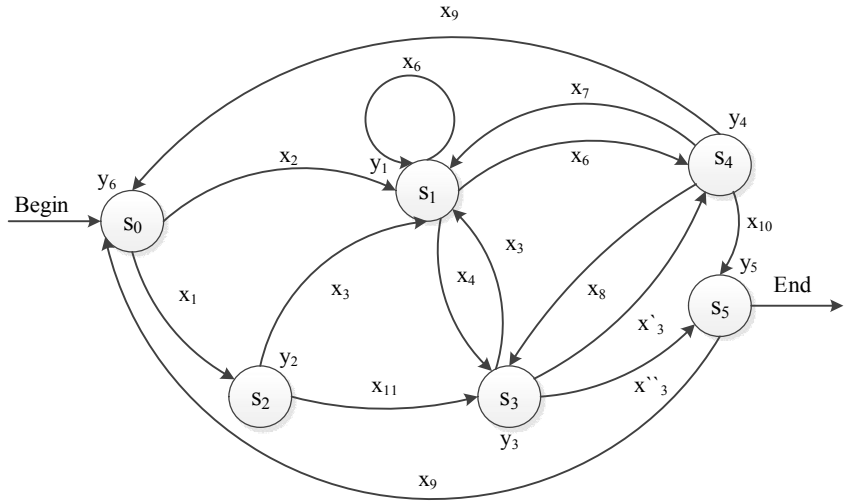


Fig. 2. Moore automatic machine of Boyd cycle with the usage of ontology

The area of military technology is characterized by the absence of regions regulatory established definitions and strict classification of technologies. Military technologies are constantly developing; this is reflected in expanding and changing conceptual scheme for system. Elaboration of formal ontology, which includes in axiomatic component, for such domain conditions an extremely difficult problem. Military technology is a complex-structured sphere that includes abstract, general

notions as well as applied terminology, which contains concepts in the specific implementation of military technologies. Four main levels of the ontological model of military technologies structure can be established. The first level is the anthology of knowledge presentation. The goal of the first level is to create a language for specification of lower level ontology. Since sphere of military technologies are a subclass of technology, relevant ontology of the upper level was introduced. Upper level ontology can be used as a basis for elaboration of ontologies of different domains. It describes main concepts in the sphere of technologies such as "technology", "knowledge", "technofact", "production technology", "dual technology", "product", etc. The third level includes ontology of military technologies domain. Main concepts of the military technologies domain are: "military technology", "weaponization technology", "weapon production technology", "basic military technology", "critical military technology", "list of basic and critical military technologies", "basic military technologies development program". Applied ontology of military technology, the forth level constituent, describes the set of military technology implementations. It contains specific information – concepts and relations, which reveal peculiarities of certain types of weapon and military technology (laser weapons, reactive armor, all-hypersonic platform, navigation systems, etc.) [6-13]. Domain terms in a given case are: combat vehicles, cannons, cannon artillery projectile, etc. The connections between terms are: "has a projectile", "has a canoon", etc. (Fig. 3).

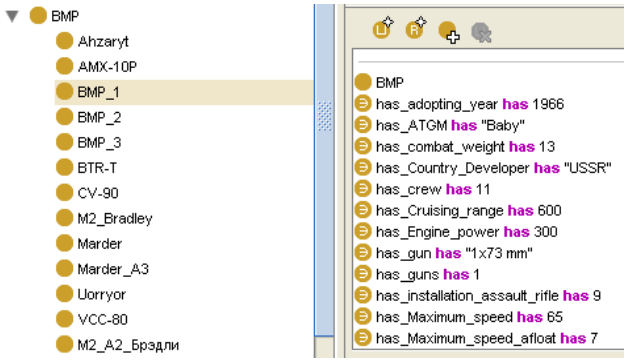


Fig. 3. Ontology fragment and determination of its individual elements

In order to increase the efficiency of possible decisions in ontology, the knowledge of experts is presented (generals and colonels of Ukrainian Ground Forces) concerning the behavior in certain situations by means of descriptive logic (DL). For example, an expert rule "to bring down fire of our artillery to man-portable air defense system of an enemy during landing of our troops from helicopter on the territory x if the distance to man-portable air defense system of the enemy is less or equal to y" on the DL language in our ontology is presented in the following way: (Landing (Troops, ?x)) ∧ (Location (man-portable air defense system of an enemy, ?x) ≤ ?y) → Bring down fire (our Artillery, man-portable air defense system of an enemy). On the basis of the approach developed, a DSS was elaborated. We experimentally proved the efficiency of our DSS, which allowed to reduce the time

spent by the Army leaders on the operational planning and task communication to subordinates; optimization of organizational and staff structures, joints, units and subunits of the Ukrainian Ground Forces [19-26]; improvement of operative and combat training of the Ukrainian Ground Forces [27-40].

3 Conclusions

This work contains a solution of the important scientifically and practically task of developing methods and tools for building support decision systems in a competitive environment (military area) using ontological approach. The efficiency of such systems which is achieved through the use of mathematical and software developed based on the use of ontology in these systems and adapting to specific problems of the ontology domain. The expediency of development of mathematical models, methods and tools for support decision in a competitive environment is based on Boyd loops using ontological approach in those subject areas where knowledge is explicit. One of such subject areas is the military sphere. The models based on a hinged Boyd-Moore automaton. State machine loops are the stages Boyd and filling processes, ontology editing and a search of relevant knowledge ontology. Possible transitions between the states of the machine and transfer parameters between them were determined. To simulate the process of decision support in a competitive environment mathematical software and methods of using domain ontology in four stages loops OODA (observation, orientation, decision, action) were developed. Thus, for the military at an "Observations" stage intelligence domain ontology determining of the strengths and weaknesses of the enemy. At the stage of "orientation", ontological data are used for simulation of possible realizations of the battle and for optimal positioning of the forces. For "Resolution" stage, target distribution was developed using a method based on genetic algorithms, which helped to reduce the computational complexity of finding effective distribution targets. Ontology expertise based on descriptive logic is presented to increase the effectiveness of possible solutions.

References

1. Hatzi, O., Vrakas, D., Bassiliades, N., Anagnostopoulos, D., Vlahavas, I.: The PORSCE II Framework: Using AI Planning for Automated Semantic Web Service Composition. In: The Knowledge Engineering Review, Cambridge University Press, 2(3), 1-24 (2010)
2. Gruber, T.: A translation approach to portable ontologies. In: Knowledge Acquisition, 5/2(2), 199-220 (1993)
3. Link Grammar, Carnegie Mellon University. Available: http://bobo.link.cs.cmu.edu/link.
4. Guarino,N.: Formal Ontology, Conceptual Analysis and Knowledge Representation. In: International Journal of Human-Computer Studies, 43(5-6), 625-640 (1995)
5. Sowa, J.: Conceptual Graphs as a universal knowledge representation. In: Semantic Networks in Artificial Intelligence, Computers & Mathematics with Applications, 23(2–5), Part 1, 75-95 (1992)

6. Montes-y-Gómez, M., Gelbukh, A., López-López, A.: Comparison of Conceptual Graphs. In: Lecture Notes in Artificial Intelligence, 1793, http://ccc.inaoep.mx/~mmontesg/publicaciones/ 2000/ComparisonCG.
7. Bulskov, H., Knappe, R., Andreasen, R.: On Querying Ontologies and Databases. In: FQAS, 191-202 (2004)
8. Callı, A., Gottlob, G., Pieris, A.: Advanced processing for ontological queries. In: Proc. of the 36th International Conference on Very Large Databases, 554-565 (2010)
9. Fensli Rune, Jan Pettersen Nytun Master: An Ontology-Based Decision Support System for Interventions based on Monitoring Medical Conditions on Patients in Hospital Wards By Tian Zhao Supervisors. In: Proceedings of the International Conference on Information and Communication Technology, Spring, Faculty of Engineering and Science University of Agder Grimstad, June 2 (2014).
10. Ugon, A., Sedki, K., Kotti, A., Seroussi, B., Philippe, C., Ganascia, J., Garda, P., Bouaud, J., Pinna, A.: Decision System Integrating Preferences to Support Sleep Staging. In: Studies in health technology and informatics, 228, 514-518 (2016)
11. Galopin, A., Bouaud, J., Pereira, S., Seroussi, B.: An Ontology-Based Clinical Decision Support System for the Management of Patients with Multiple Chronic Disorders. In: Studies in health technology and informatics, 216, 275-279 (2015).
12. M. Rospocher, L. Serafini, "An Ontological Framework for Decision Support Joint", in Proceedings of the International Semantic Technology Conference, 2012, pp 239-254.
13. Rospocher, M., Serafini, L.: Ontology-centric decision support. In: Proc. of the Inter. Conf. on Semantic Technologies Meet Recommender Systems & Big Data, 919, 61-72 (2012)
14. Lytvyn, V., Vysotska, V., Chyrun, L., Chyrun, L.: Distance Learning Method for Modern Youth Promotion and Involvement in Independent Scientific Researches. . In: International Conference on Data Stream Mining & Processing, 269-274 (2016)
15. Lytvyn, V., Vysotska, V., Burov, Y., Veres, O., Rishnyak, I.: The Contextual Search Method Based on Domain Thesaurus. In: Advances in Intelligent Systems and Computing, 689, 310-319 (2017)
16. Vysotska, V.: Linguistic Analysis of Textual Commercial Content for Information Resources Processing. In: Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET, 709-713 (2016)
17. Su, J., Vysotska, V., Sachenko, A., Lytvyn, V., Burov, Y.: Information resources processing using linguistic analysis of textual content. In: Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 573-578 (2017)
18. Lytvyn, V., Vysotska, V., Veres, O., Rishnyak, I., Rishnyak, H.: Content Linguistic Analysis Methods for Textual Documents Classification. In: International Conference on Computer Science and Information Technologies, Lviv, Ukraine, 190-192 (2016)
19. Mykich, K., Burov, Y.: Research of uncertainties in situational awareness systems and methods of their processing. In: EasternEuropean Journal of Enterprise Technologies, 1(79), 19-26 (2016)
20. Vysotska, V., Rishnyak, I., Chyrun, L.: Analysis and evaluation of risks in electronic commerce. In: CAD Systems in Microelectronics, CADSM '07, 9th International Conference, Lviv, 332-333 (2007)
21. Vysotska, V., Chyrun, L.: Analysis features of information resources processing. In: Computer Science and Information Technologies, 124-128 (2015)
22. Lytvyn, V., Vysotska, V.: Designing architecture of electronic content commerce system. In: Computer Science and Information Technologies, Proc. of the X–th Int. Conf. CSIT, 115-119 (2015)

23. Khomytska, I., Teslyuk, V.: Modelling of Phonostatistical Structures of English Backlingual Phoneme Group in Style System. In: Proceedings of the 14th International Conference on CADMS, 324-327 (2017)
24. Khomytska, I., Teslyuk, V.: Modelling of Phonostatistical Structures of the Colloquial and Newspaper Styles in English Sonorant Phoneme Group. In: XIIth Conference on CSIT, Lviv, 67-70 (2017)
25. Teslyuk, V., Beregovskyi, V., Denysyuk, P., Teslyuk, T., Lozynskyi, A.: Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System. In: Intern. Journal of Intelligent Systems and Applications, 10(1), 1-8 (2018)
26. Bondariev, A., Kiselychnyk, M., Nadobko, O., Nedostup, L., Chyrun, L., Shestakevych, T.: The software complex development for modeling and optimizing of processes of radio-engineering equipment quality providing at the stage of manufacture. In: TCSET, (2012)
27. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Germany: LAP LAMBERT Academic Publishing (2016).
28. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
29. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
30. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
31. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
32. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
33. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
34. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
35. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
36. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 100–109 (2017)
37. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)
38. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
39. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
40. Hamon, T.: Biomedical text mining. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf (2017)

Methods of Information Resources Processing in Virtual Library

Bohdan Rusyn[1], Victoria Vysotska[2], Liubomyr Pohreliuk[3]

[1]Kazimierz Pulaski University of Technology and Humanities in Radom, Poland
[2]Lviv Polytechnic National University, Lviv, Ukraine
[1,3]Karpenko Physico-Mechanical Institute of the NAS Ukraine

[1] rusyn@ipm.lviv.ua,
[2]Victoria.A.Vysotska@lpnu.ua,
[3]liubomyr@inoxoft.com

Abstract. The thesis discusses the development of unified methods and software tools for processing information resources in Virtual Library as example the electronic content commerce systems (ECCS). A model of Virtual Library Systems (VLS) is proposed. The models of information resource processing in ECCS are proposed. Architecture and models of VLS are built. A new approach to application and implementation of business processes is formulated for the construction of VLS. A complex method of content creation, the operational method of content management and complex method of content support are developed. Design and implementation methods of VLS are based on online newspapers, which reflect the results of theoretical research, are developed. From the perspective of a systemic approach, the principles of applying information resources processing in electronic content commerce systems for content lifecycle implementation made the development of methods for the content formation, management and support possible. An integrated method of content formation for the time and resources reduction of content production is developed. A method of content management for the time and resources reduction of content sales was created, which makes it possible to implement content management subsystem. A method of content support for the time and resource reduction of the target audience analysis in VLS is implemented, which makes it possible to develop a content support subsystem.

Keywords: content, commercial content, information resource, business-process, content management system, content lifecycle, Internetnewspaper, electronic content commerce system, Virtual Library.

1 Introduction

Rapid development of the Internet contributes to the increase of needs for the efficient data of the production / strategic nature and implementation of new forms of information services through modern information technologies (IT) of e-commerce.

Documented information prepared in accordance with users needs is a commercial content. Today e-commerce is a reality and a promising business process. Internet is the business environment, and commercial content is a commodity with the highest demand and selling rate. It is also the main object of the processes of electronic content commerce. Commercial content can be immediately ordered, paid and got on-line as a commodity. The entire spectrum of commercial content is sold via the Internet - scientific and publicistic articles, music, books, movies, pictures, software etc. Well-known corporations that implement electronic content commerce are Google through Google Play Market, Apple - Apple Store, Amazon - Amazon.com. Most of the decisions and researches are conducted at the level of specific projects. Systems of electronic content commerce (SECC) are built on the closed principleasnon-recurrent projects. Modern SECC are focused on the realization of commercial content that is made outside of the system. Design, development, implementation and maintenance of SECC are impossible without the use of modern methods and information technologies of formation, management and maintenance of commercial content.

2 Relevance of the paper

Development of the technology of information resources processing is important in view of such factors as lack of theoretical grounding of methods of study of commercial content flows and the need for unification of software processing methods of information resources in ECCS. A practical factor of the processing of information resources in VLS is related with the solution of problems of formation, management and support of growing volumes of commercial content in the Internet, rapid development of e-business, widely spreaded availability of the Internet, the expansion of the set of information products and services, and increasing of a demand for commercial content. Principles and IT of electronic content commerce are used while creating on-line stores (selling of e-Books, Software, video, music, movies, picture), on-line systems (newspapers, magazines, distance education, publishing) and off-line selling of content (copywriting services, Marketing Services Shop, RSS Subscription Extension), cloud storage and cloud computing. The world's leading producers of means of processing of information resources as Apple, Google, Intel, Microsoft, and Amazon are working in this area [2, 4-7].

A theoretical factor of information resources processing in ECCS is connected with the development of IT processing of commercial content. In scientific studies of D. Lande, V. Furashev, S. Braychevsky, A. Grigoriev mathematical models of electronic processing of information flows are investigated and developed [1, 10-11]. G. Zipf proposed an empirical law of distribution of word frequencies in natural language text content for its analysis. In the works of B. Boiko, S. McKeever, A. Rockley models of the life cycle of content are developed [1-21]. The methodology of content analysis for processing textual data sets was initiated and developed by M. Weber, J. Kaiser, B. Glaser, A. Strauss, H. Lasswell, O. Holsti, Ivanov, M. Soroka, A. Fedorchuk. In the works of V. Korneev, A. Gareev, S. Vasyutin, V. Reich were proposed methods of intellectual processing of text information. EMC, IBM,

Microsoft Alfresco, Open Text, Oracle and SAP have developed specification of Content Management Interoperability Services based on Web-services interface to ensure interoperability of electronic content commerce system management [4-7]. From the scientific point of view, this segment of IT is not investigated enough. Each individual project is implemented almost from the very beginning, in fact, based on the personal ideas and solutions [22-31]. In literature, very few significant theoretical studies, research findings, recommendations for the design of VLS and processing of information results in such systems are highlighted. Appeared a need to analyze, to generalize and to justify existing approaches to implementation of e-commerce and building of VLS. The actual problem of the creation of technological products complex is based on the theoretical study of methods, models and principles of processing information resources in VLS, based on the principle of open systems that allow to manage the process of increase in sales of commercial content. Analysis of the factors enables us to infer the existence of a contradiction between the active development and extension of IT and VLS on the one hand, and the relatively small amount of research on this subject and their locality on the other. This contradiction raises the problem of containment of innovation development in the segment of electronic content commerce through creation and introduction of appropriate new advanced IT that affects negatively the growth of this market [32-45]. Within this problem there is an urgent task of developing scientifically based methods of processing information resources of electronic content commerce, and building process on the basis of software for the creation, dissemination and sustainability of VLS. In this paper a study to identify patterns, characteristics and dependencies in information resources processing in VLS was carried out.

3 The Main Content of Work

The basic terms and concepts had been defined and concretized. The following of them had been used in the work. The *content* is the totality of all data (commercial, service, extra, etc.) that implement a certain set of meta-models (a model that describes the structure and principles of a particular model) and the models of copies concentrated among information system. The *commercial content* is a part of the general content, which is the subject of the purchase, the user's use and owner's profit; textual, visual or audio content as part of the user's experience according to the information resources (text, images, audio, video and software). The content control is control functions for receiving, analyzing, saving, searching and spreading of the content. The information resource is an object of the means' action and information technology; set of documents in the information systems (libraries, archives, data banks, etc.).

The information product is documented information prepared and designed to meet the needs of users.

The e-commerce is a field of digital economy and of e-business, including all financial and commercial transactions over computer networks and business processes associated with conducting these transactions. E-content commerce is a field of e-commerce, where the commercial content is an object of financial and commercial

transactions and business processes. The system of e-content commerce is a system of processing of commercial content and related information, human, technical, organizational and financial resources, to support and distribute commercial content. The content lifecycle is a multi-complex process that takes place in the content control via the various stages / phases of the publication with a set of properties such as collaboration, records' control, digital asset and versions that are supported by various technologies. Large torrents and volumes of different content are in VLS. Most of these content's torrents are made up of easily formalized and automated procedures and commercial content. But there is no general approach to the process of modeling, design, development and implementation of VLS. The formal description of the VLS is presented as

$$Y = \langle X, Q, C, V, H, Z, T, \delta \rangle, \quad (1)$$

where $X = \{x_1, x_2 \ldots, x_{n_X}\}$ is a set of content from various sources (information re-resources, authors, moderators, editors, visitors, journalists, users, administrators, analysts), $Q = \{q_1, q_2 \ldots, q_{n_Q}\}$ is set of users' information requests, $C = \{c_1, c_2 \ldots, c_{n_C}\}$ is a content set, $V = \{v_1, v_2 \ldots, v_{n_V}\}$ is a set of conditions content maintenance and external influences on the system environment, $H = \{h_1, h_2 \ldots, h_{n_H}\}$ is a set of processing content's conditions, $Z = \{z_1, z_2 \ldots, z_{n_Z}\}$ is set of information resource's components, $T = \{t_1, t_2 \ldots, t_{n_T}\}$ is time of transaction processing content, $Y = \{y_1, y_2 \ldots y_{n_Y}\}$ is an outgoing characteristics' totality of the system, δ is an operator which form the statistics' analysis of VLS's functioning. The process which works up information resources (1) is described by the operator $y_j(t_{p+1}) = \delta(x_i, q_d, c_r, v_l, h_k, t_p, z_w)$.

The value $y_j = \{y_{1j}, y_{2j}, \ldots, y_{gj}\}$ is a totality of data over a specified period of time, where y_1 is number of visits, y_2 is average time of information resource's attendance (min: c), y_3 is a rate of refusals (%), y_4 is an achieved goal of a search, y_5 is content's dynamic (%), y_6 is the total number of viewed pages, y_7 is number of viewed pages per visit, y_8 are new visits, y_9 are absolute unique visitors, y_{10} is a traffic's source in % and so on. The impact of the values x_i, q_d, c_r, v_l, h_k on the values z_w and y_j as a result of the e-commerce's content are unknown and unexplored. Connections between the input data, content, input data and the processing of information resources in the system are undisclosed. This justifies a goal, an actuality, an expediency and a research's areas.

4 Information Resources Processing

The main stages of the process of information resources' elaboration in VLS are

formation, control and maintenance of commercial content, with the following links: *content → content's formation → database → content's control→ informational resource or user's request → content's control→ informational resource → content's maintenance → database.* Then from (1) $\delta : X \to Y$ conveyed functions' superposition $\delta = \gamma \circ \beta \circ \alpha$, where α is an operator of commercial content's formation, β is an operator of commercial content's control, γ is an operator of commercial content's maintenance. The VLS is presented as

$$Y = \langle X, Q, H, C, V, Z, T, \alpha, \beta, \gamma \rangle . \tag{2}$$

The operator of commercial content's formation α is a commercial content's reflection c_r into new state c_{r+1}, that differs from the previous due to emergence of a new piece of content Δc which complements the previous state $c_{r+1} = c_r + \Delta c$, then $\alpha : (c_r, t_p, X, u_f) \to (c_{r+1}, t_{p+1})$, where $u_f = \{u_{1f}, u_{2f}, \ldots, u_{n_U f}\}$ is set of formation of commercial content's conditions c_r as

$$c_r = \left\{ \bigcup_i^{n_X} x_i \left| \begin{array}{l} \forall x_i \in X_{u_f}, x_i \notin X_{\overline{u_f}}, \exists u_f \in U_{x_i}, u_f \notin U_{\overline{x_i}}, \\ X = X_{u_f} \cup X_{\overline{u_f}}, U = U_{x_i} \cup U_{\overline{x_i}}, f = \overline{1, n_U} \end{array} \right. \right\}, \tag{3}$$

where the set of conditions u_f commercial content's formation c_r is defined as

$$u_f = \left\{ \bigcup_j^{k} u_{jf} \left| \begin{array}{l} \forall u_{jf} \in U_{x_i}, \exists x_i \in X_{u_f}, u_{jf} \notin U_{\overline{x_i}}, \\ U = U_{x_i} \cup U_{\overline{x_i}}, X_{u_f} \subseteq X, f = \overline{1, n_U}, i = \overline{1, m} \end{array} \right. \right\}. \tag{4}$$

The operator of commercial content's control β is a reflection of commercial content c_r into new state c'_r, which is different from the previous state due to values of the defining parameters $h_k \to h'_k$ (actuality, completeness, relevance, authenticity, trustworthiness) that satisfy predefined requirements

$$\beta : (q_d, z_w, c_r, h_k, u_M, t_p) \to (c'_r, h'_k, z_{w+1}, t_{p+1}), \tag{5}$$

where $q_d \in Q$, $h_k \in H$, $h_k = \{h_{1k}(c_r, q_d), \ldots, h_{n_H k}(c_r, q_d)\}$ is set of conditions of commercial content's control as

$$z_w = \left\{ \bigcup_{r=1}^{n_C} c_r \left| \begin{array}{l} \forall c_r \in C_{q_d}, \exists q_d \in Q, \exists h_k \in H_{c_r}, c_r \notin C_{\overline{q_d}}, \\ h_k \notin H_{\overline{c_r}}, C = C_{q_d} \cup C_{\overline{q_d}}, H = H_{c_r} \cup H_{\overline{c_r}}, \\ d = \overline{1, n_Q}, k = \overline{1, n_H} \end{array} \right. \right\}, \tag{6}$$

where the set of defining parameters' values form as $h'_k = h_k + \Delta h$. The operator of commercial content's maintenance γ is a commercial content reflection c_r in the collection of values y_i, which is formed as result of the analysis, monitoring, evaluation of user's interaction, searching engines and other information resources that are the basis for making decisions about development and commercial content's control

$$\gamma : (c_r, q_d, v_l, h_k, z_w, u_S, t_p) \to y_i, \tag{7}$$

where $v_l = \{v_{1l}(q_i, h_k, c_r, z_w, t_p), \ldots, v_{n_V l}(q_i, h_k, c_r, z_w, t_p)\}$ is set of conditions of content's maintenance and impact of the environment on the system. Outgoing data is implemented in

$$y_j = \left\{ \bigcup_{l}^{n_V} v_l \left| \begin{array}{l} \forall v_l \in V_{q_d} \cup V_{z_w}, \exists q_d \in Q, \exists z_w \in Z, \exists h_k \in H_{c_r}, \\ v_l \notin V_{\overline{q_d}}, v_l \notin V_{\overline{z_w}}, V_{q_d} \subset V, V_{z_w} \subset V, d = \overline{1, n_Q}, \\ w = \overline{1, n_Z}, r = \overline{1, n_C}, k = \overline{1, n_H} \end{array} \right. \right\}. \tag{8}$$

The process of commercial content's formation for an information resource provides a mapping of the input data from different sources to the set of formed commercial content and saved in an appropriate database in the VLS as $S(x_i) \to x_i \to X \to \alpha(u_f, x_i, t_p) \to c_r \to C \to D(C)$, where $S(x_i)$ is a data source, $D(C)$ is database of commercial content. The commercial content's formation $\alpha : X \to C$ is presented by superposition of functions

$$\alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_0, \tag{9}$$

$$\alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_1, \tag{10}$$

where α_0 is an operator of commercial content's creating; α_1 is an operator collecting content from multiple sources; α_2 is an operator which identifies duplication of commercial content; α_3 is an operator commercial content's formation; α_4 is an operator which identifies key words and commercial content's concepts; α_5 is an operator of commercial content's automatic categorization; α_6 is an operator which forms commercial content's digests; α_7 is operator of commercial content's selective distribution. The process of commercial content's formation is presented as

$$\alpha = \langle X, T, U, C, \alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7 \rangle. \tag{11}$$

1. The operator of the commercial content's establishment is mapping of input data from various sources of information into commercial content that is different

from the previous state of the commercial content due to its actuality as $\alpha_0 : (X, U_C, T) \to C_0$.

2. The operator of the commercial content's collecting is reflection of input data from the authors or systems' moderators in commercial content that is different from the previous state of the commercial content due to its trustworthiness and actuality as $\alpha_1 : (X, U_G, T) \to C_0$.

3. The operator which identifies commercial content's duplication is a reflection of a commercial content into a new state that is different from the previous state according to its uniqueness as $\alpha_2 : (C_0, T, U_B) \to C_1$.

4. The operator of commercial content's formatting is display content in a new state that is different from the previous state according to its format of presentation as $\alpha_3 : (C_1, U_{FR}, T) \to C_2$.

5. The operator which identifies commercial content's keywords is a commercial content's reflection into a new state that is different from the previous state due to the presence of the set of keywords that describe the general content as $\alpha_4 : (C_2, U_K, T) \to C_3$.

6. The operator which categorizes content is a content's reflection into a new state due to it's validation, which is different from previous state due to it's belonging to the set of thematic content as $\alpha_5 : (C_3, U_{CT}, T) \to C_4$.

7. The operator which forms commercial content's digests is a commercial content's reflection into a new state that is different from the previous state due to the emergence of a new piece of content as a summary of its complement previous state as $\alpha_6 : (C_4, U_D, T) \to C_5$.

8. The operator of commercial content's selective distribution is a commercial content's reflection into a new state that is different from the previous state due to its purpose and spread among the target audience as $\alpha_7 : (C_5, U_{Ds}, T) \to C_6$.

The set of operators $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$ is adequate in the process of commercial content. The process of commercial content is presented by the following scheme of links: $User(q_d) \to q_d \to Q \to H(c_r, q_d) \to \beta(q_d, c_r, h_k, t_p) \to z_w \to User(z_w)$, where $User(q_d)$ is a formation of user's request; $User(z_w)$ is browsing by user's answers to a request q_d. The operator of commercial content's control $\beta : C \to Z$ is presented as superposition of functions $\beta = \beta_4 \circ \beta_3 \circ \beta_2 \circ \beta_1$, where β_1 is an operator of editing and modification of commercial content; β_2 is an operator which determines the weight of the block of commercial content; β_3 is an operator which form values of defining parameters of commercial content's control; β_4 is an operator which form and present information resource pages. The commercial content management is presented as

$$\beta = \langle C, Q, H, U, T, Z, \beta_1, \beta_2, \beta_3, \beta_4 \rangle . \quad (12)$$

1. The operator of the editing and modification of commercial content is

presented as $\beta_1 : (c_r, h_k, u_l, t_p) \to c'_r$.

2. The determination operator of the block weight and formation the base search images of commercial content is presented as $\beta_2\ (c'_r, y_j, u_l, t_p) \to c''_r$.

3. The formation operator of the determinant meanings parameters of control is presented as $\beta_3 : (c''_r, h_k, u_l, t_p) \to h'_k$.

4. The formation and presentation operator of information resource page is presented as $\beta_4\ (c''_r, h'_k, z_w, q_d, t_p) \to z_{w+1}$, where $h_k \in H$, $h_k, = \{h_{1k}, h_{2k}, ..., h_{mk}\}$ is the set of process parameters that control commercial traditional content (h_{1k} is actuality, h_{2k} is relevance, h_{3k} is completeness, h_{4k} is authenticity, h_{5k} is authenticity of commercial content); $u_l \in U$, $u_l = \{u_{1l}, u_{2l}, ..., u_{nl}\}$ is multiple criteria process to control commercial content (u_{1l} is the coefficient of the block location in the commercial content, u_{2l} is the coefficient of keywords in the block , u_{3l} is coefficient of the key words statistical importance, u_{4l} is the coefficient of keywords from the user request, the coefficient of the keywords volume from the request). The process of commercial content maintenance is presented as scheme of links: $User(q_d, z_w) \to q_d \to z_w \to V(q_d, z_w, t_p) \to \gamma(v_l, h_{k,} c_r, z_w, t_p) \to y_j \to Moderator(y_j)$.

Support of commercial content $\gamma : Z \to Y$ is represented by a superposition of functions $\gamma = \gamma_8 \circ \gamma_6 \circ \gamma_5 \circ \gamma_3 \circ \gamma_1$, or $\gamma = \gamma_8 \circ \gamma_7 \circ \gamma_5 \circ \gamma_4 \circ \gamma_2$, where γ_1 is the operator of formation digital flows portraits of commercial content , γ_2 is the operator of formation digital portraits of regular users, γ_3 is the operator of identification of thematic subjects in the plural of new commercial content, γ_4 is the operator of identification of commercial content thematic subjects with a set of user requests, γ_5 is the operator of tabulation of the commercial content relations, γ_6 is the operator of calculating the ratings of commercial content, γ_7 is the operator of calculating the ratings of regular users, γ_8 is operator of the statistical analysis system functioning.

The process of commercial content support is presented as

$$\gamma = \langle Q, C, H, V, T, Z, Y, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7, \gamma_8 \rangle, \tag{13}$$

1. The operator of formation digital flows portraits of content. It is a mapping of the set of relevant commercial content in a variety of settings, which describe the thematic needs of the target audience according to certain criteria that is defined by the moderators as $\gamma_1 : (V_{Pc}, C, H, Q, T) \to Y_{Pc}$.

2. The operator of formation portraits of regular users is the mapping of the relevant commercial content set in a variety of settings which describe the thematic needs of the target audience according to certain criteria that is defined by moderators as $\gamma_2 : (V_{Pq}, Q, H, Z, T) \to Y_{Pq}$.

3. The operator of identification of thematic subjects in the plural of new

commercial content which is mapping of new commercial content set from a variety of sources of information in the set of keywords for new rubric of commercial content which describe a topic sentence of these reliable sources according to certain criteria that is defined by the moderators as $\gamma_3 : (C, H, X, V_T, T) \to Y_T$.

4. The operator of identification of content thematic subjects with a set of user requests is the mapping of multiple user requests to the set of keywords for the new rubric of content which describe the thematic needs of registered users according to certain criteria that is defined by the moderators as $\gamma_4 : (C, H, Q, V_T, T) \to Y_T$.

5. The operator of tabulation of the commercial content relations by keywords and frequency of visits is the mapping of commercial content in a new state, which is different from the previous large number of links content based on criteria such as thematic, the relevance factor rating, sequence and frequency of viewing, popularity, actuality, authorship as $\gamma_5 : (C, V_c, T) \to Y_C$.

6. The operator of calculating the ratings of content is the mapping of commercial content to a new state which is different from the previous state of commercial content by the emergence of new content in the form of ratings on certain criteria, that complements the previous state as $\gamma_6 : (C, Q, H, Y_C, V_{Rc}, T, \theta, \vartheta) \to Y_{Rc}$.

7. The operator of calculating the ratings of regular users – the mapping of the set permanent portraits of classified users in a new state , which is different from the previous condition of commercial content by the emergence of a new part of the characteristics of these users in the form of ratings on certain criteria, that complements the previous state as $\gamma_7 : (C, Q, H, Y_C, V_{Rm}, T) \to Y_{Rm}$.

8. The operator of the statistical analysis of system functioning is the mapping of statistic system functioning in a collection of values, which create as result of analysis, monitoring, evaluation of user interaction, search engines and other information resources, which is the basis of making decisions regarding to the content creation and management as $\gamma_8 : (Y_P, Y_T, Y_C, Y_R, Z, H, V, T) \to Y$.

The subsystem of content formation is implemented as a content-monitoring complex for content gathering from different sources of data which provides a content database creation according to the information needs of users. As a gathering and primary processing result of content is reduced to a single format, classified according to the specified categories. And he is credited descriptors with keywords. This facilitates the process implementation of content management. Tasks of Web content management subsystem are: database formation, rotation and providing access to it; the operational and retrospective databases formation; the user experience personalization; personal user queries and sources storing; operation statistics analysis; search providing in database; initial forms generation on information resources; information interaction with other databases; the an information resource formation. Content management subsystem is implemented through caching (representation module generates a page once; then it is several times faster loaded from the cache, which is updated automatically after a certain period of time or when making changes to specific sections of an information resource, or manually by administrator command) or information blocks formation (blocks conservation in the information resources editing stage and page collection from these blocks at the user

request of the relevant page). Content support subsystem provides information portraits formation, thematic storyline identification in content flows, the content relationship tables building, content rankings calculation, new events identification in their content flows, their tracking and clustering. Analysis of commercial support content helps identify causes of the formation of the target audience for a set of characteristics of functioning of VLS. By adjusting the a set of thematic commercial content, its uniqueness, efficiency of its formation and adequate management according the individual needs of the regular user, you can to model the verge of the target social audience and the number of unique visitors from search systems.

5 Conclusion

The paper is solved the actual scientific problem of methods and means research and development for commercial Web content processing in e-business systems by using the developed mathematical software for the appropriate systems creation, which made it possible maintain the life cycle of commercial content on the level of developer (the time and costs reducing for development, quality improvement through the use of proven solutions).

References

1. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2016).
2. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
3. CM Lifecycle Poster, Content Management Professionals, Access:http://www.cmprosold.org/resources/poster/ (2010)
4. Content Management Interoperability Services, Version V. 0.5, Hopkinton: EMC (2008)
5. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
6. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
7. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
8. Hackos, J.: Content Management for Dynamic Web Delivery. Hoboken: Wiley (2002)
9. Halvorson, K.: Content Strategy for the Web. Reading: New Riders Press (2009)
10. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 47–55 (2017)
11. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)
12. McGovern, G., Norton, R.: Content Critical. Upper Saddle River: FT Press (2001)

13. McKeever, S.: Understanding Web content management systems: evolution, lifecycle and market. In: Industrial Management & Data Systems (MCB UP), 103(9), 686–692 (2003)
14. Nakano, R.: Web content management: a collaborative approach. Boston: Addison Wesley Prof. (2002)
15. Papka, R.: On-line News Event Detection, Clustering, and Tracking: thesis for the degree doctor of philosophy. Amherst: Massachusetts University (1999)
16. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
17. Rockley, A.: Managing Enterprise Content: A Unified Content Strategy. New Riders Press (2002)
18. Stone, W. Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! In: IEEE Antennas and Propagation, 45(4) 47-49 (2003)
19. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
20. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
21. The Content Management Possibilities Poster. http://metatorial.com/pagea.asp?id=poster.
22. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
23. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. . In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 100–109 (2017)
24. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)
25. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
26. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
27. Hamon, T.: Biomedical text mining. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf (2017)
28. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)
29. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
30. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)

31. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)
32. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
33. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf. (2017)
34. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
35. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)
36. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
37. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 133–134 (2017)
38. Kirkin, S., Melnyk, K.: Intelligent data processing in creating targeted advertising. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 131–132 (2017)
39. Hordienko, H., Ilchenko, M.: Development and computerization of an English term system in the fields of drilling and drilling rigs. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 129–130 (2017)
40. Gorbachov, V., Cherednichenko, O.: Improving communication in enterprise solutions: challenges and opportunities. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 127–128 (2017)
41. Didusov, V., Kochueva, Z.: Statistical methods usage of descriptive statistics in corpus linguistic. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 125–126 (2017)
42. Verbinenko, Yu.: Discursive units in scientific texts. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 120–123 (2017)
43. Titova, V., Gnatchuk, I.: Evaluation ofa formalized model for classification of emergency situations. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 110–119 (2017)
44. Kotov, M.: NLP resources for a rare language morphological analyzer: danish case. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 31–36 (2017)
45. Kuprianov, Ye.: Semantic state superpositions and their treatment in virtual lexicographic laboratory for spanish language dictionary. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 37–46 (2017)

SEO Technology for Web-Resource Processing

Lyubomyr Chyrun[1], Victoria Vysotska[2], Liliya Chyrun[3],
Aleksandr Gozhyj[4], Iryna Kalinina[5]

Lviv Polytechnic National University, Lviv, Ukraine
Petro Mohyla Black Sea National University, Mykolayiv, Ukraine

[1]Lyubomyr.V.Chyrun@lpnu.ua,
[2]Victoria.A.Vysotska@lpnu.ua,
[3]lchirun21@gmail.com,
[4]alex.gozhyj@gmail.com,
[5]irina.kalinina@gmail.com

Abstract. The thesis to the methods of elaborating Web-resources in electronic commercial content systems is devoted. The methods of formation and maintenance of commercial Web-content were developed and the method of the management of commercial Web-content based on SEO technology are improved.

Keywords: content, commercial Web-content, Web-resource, business-process, content management system, content lifecycle, electronic content commerce system, SEO technology.

1 Introduction

The purpose of the project is the implementation of standardized testing methods and software processing approbation of Web-resources in electronic content commerce system (ECCS) [4-7]. The formation of ECCS overall architecture promotes the generalization of ECCS information resources technique through the stages of formation, management and maintenance of Web-content in order to reduce the time while constructing e-business common systems. The implementation of ECCS contributes to the reduction of time while production of own commercial Web-content, analysis of external Web-content from other sources [1-2, 10-11], commercial content lifecycle dynamic analysis [3, 8-9, 12-18], ECCS functioning statistical analysis [2, 19-27], statistical analysis of information resources user activities in ECCS, increasing of information resources target audience and extension of ECCS functional capabilities [28-35].

2 Recent research and publications analysis

The purpose of ECCS is formation, management and support of commercial Web-

content on principles of Web-resources processing. ECCS is designed to create common functional requirements and standardized specifications concerning development through processing stages optimization of the Web-resources in similar systems [1-7, 35-45].

The list of tasks is performed by ECCS [1-27].

- Formation of Web-content (collecting data from various sources and their formation, identifying keywords and duplication, digest formation, categorization and content selective distribution, content creation, maintaining Web-content, creation of filtering content rules).
- Commercial Web-content management (formation/ rotation of databases and access to them, subscribing on thematic content, content distribution, individualization of users work, storing of users' requests and sources, keeping operation statistics; search providing; generation of output forms; information interaction with databases; formation of information resource, formation of comments and content feedbacks, voting on content).
- Commercial Web-content support (formation of content stream portraits as well as potential/constant users and target audience; identifying content thematic subjects; formation of content relationship tables; calculation of ECCS content and moderators/authors ratings; detection, monitoring, and clustering of new events in the content streams).

ECCS is used for the implementation of e-business in information service field with active usage of the Internet technology benefits. ECCS is designed to provide information services such as online newspaper, online magazine, online edition, online publishing, and online store for selling content etc. It's proposed to use ECCS in order to promote services through publishing houses, newspapers, magazines, news agencies, educational institutions, software development companies or companies which sell content without media. Types of activities where ECCS is applicable: informational (publishing, address and reference, telecommunication, provider), informational and consulting (advertising, marketing, partners reliability testing, distance education) and consulting (legal, economic, medical and other types). The spheres of application of ECCS [2, 4-7]:

- for Web-content online sales via online newspapers, online magazines, distance learning, online editions, online publishing, portals containing informative/ entertaining/children's content;
- for Web-content offline sales via such systems as copywriting services, Marketing Services Shop or RSS Subscription Extension;
- online stores for selling e-Books, video, software, music, movies, pictures, digital art, manuals, articles, certificates, forms, files etc.;
- for saving of various types of content via cloud storage or cloud computing.

ECCS is intended to solve problems related to the rapidly growth of content in the Internet or in the field of e-business activity as well as widen access to Web-resources through the Internet, active development of e-business, expanding a set of products and/or services, increasing demand for products and/or services, technologies and

means creation, and expansion of the scope of Web-resources processing methods.

3 Commercial Web-resource analysis

The ECCS is presented as $Y = \langle X, Q, H, C, V, Z, T, \alpha, \beta, \gamma \rangle$, where α is an operator of Web-content's formation, β is an operator of Web-content's control, γ is an operator of Web-content's maintenance.

1. The operator of commercial Web-content's formation α is a content's reflection c_r into new state c_{r+1}, that differs from the previous due to emergence of a new piece of content Δc which complements the previous state $c_{r+1} = c_r + \Delta c$, then $\alpha : (c_r, t_p, X, u_f) \to (c_{r+1}, t_{p+1})$, where u_f is set of formation of Web-content's conditions c_r.
2. The operator of commercial Web-content's control β is a reflection of content c_r into new state c'_r, which is different from the previous state due to values of the defining parameters $h_k \to h'_k$ (actuality, completeness, relevance, authenticity, trustworthiness) that satisfy predefined requirements
$$\beta : (q_d, z_w, c_r, h_k, u_M, t_p) \to (c'_r, h'_k, z_{w+1}, t_{p+1}),$$
where $h_k = \{h_{1k}(c_r, q_d), \ldots, h_{n_H k}(c_r, q_d)\}$, $q_d \in Q$, $h_k \in H$ is set of conditions of content's control.
3. The operator of commercial Web-content's maintenance γ is a content reflection c_r in the collection of values y_i, which is formed as result of the analysis, monitoring, evaluation of user's interaction, searching engines and other Web-resources that are the basis for making decisions about development and content's control
$$\gamma : (c_r, q_d, v_l, h_k, z_w, u_S, t_p) \to y_i,$$
where $v_l = \{v_{1l}(q_i, h_k, c_r, z_w, t_p), \ldots, v_{n_V l}(q_i, h_k, c_r, z_w, t_p)\}$ is set of conditions of content's maintenance and impact of the environment on the system.

4 Web-resources processing

The main stages of the process of Web-resources' elaboration in ECCS are formation, control and maintenance of commercial Web-content, with the following links: *content → content's formation → database → content's control→ Web-resource or user's request → content's control→ Web-resource → content's maintenance → database.*

1. Method of commercial content formation. It is a complex of measures of providing data control from different sources to create commercial content with a set of additional values (relevance, credibility, uniqueness, completeness, accuracy etc.).

The creation of commercial content is described by the operator $C_0 = \alpha_0\ (X, U_C, T)$. U_C is a set of the conditions for creation of commercial content. The task of gathering information from the sources is described by the operator $C_0 = \alpha_1\ (X, U_G, T)$, where U_G is the set of conditions of data collection from various sources . The task of identifying of doubling content matter is described by the α_2 operator in form $C_1 = \alpha_2(\alpha_0(X, U_C, T), U_B)$ and $C_1 = \alpha_2(\alpha_1(X, U_G, T), U_B)$, or $C_1 = \alpha_2(C_0, U_B)$, U_B is the set of conditions of identifying and doubling matter of commercial content.

Identifying of duplicate commercial content by content in ECCS is made with help of linguistic-statistical methods of finding common terms, lines of which form a verbal signature of commercial content (the text is unique in the factor of uniqueness $\geq 80\,\%$). The task of scanning the commercial content and the conversion to a common format in XML is described by the operator α_3 in form of

$$C_2 = \alpha_3(\alpha_2(C_0, U_B), U_{FR}), \tag{1}$$

where U_{FR} is the set of conditions of commercial content formatting. Processing of the content set C_2 to identify relevant keywords in meaning (terms) is based on the Zipf law and is reduced to the choice of words with an average frequency of occurrence (the most used words are ignored by using stop-dictionary ,and rare words are not taken to account from text messages. Identifying keywords and concepts with the use of dictionaries is determined by the operator $\alpha_4(C_2, U_K)$ in form of

$$C_3 = \alpha_4(\alpha_3(\alpha_2(C_0, U_B), U_{FR}), U_K), \tag{2}$$

when $U_K = \{U_{K1}, U_{K2}, U_{K3}, U_{K4}\}$, where U_K is the collection of conditions to identify keywords and concepts in the text, U_{K1} is the set of all terms (a term is basis of the noun, noun, word combinations of the nouns or an adjective with a noun), U_{K2} is the set of frequencies of term use in the text of content, U_{K3} is the set of coefficients of terms use that is based on the number of symbols without space (2 000 – 3 000 symbols the frequency of keywords is in the range of 4-6% , up to 2 ,000 symbols – 6-8%,more than 3,000 symbols – 2-4%), U_{K4} is the set of terms which comply with the conditions of belonging to the keywords and concepts.

The set of digests C_5 is formed depending on $C_5 = \alpha_6(C_4, U_D)$, where U_D is the set of conditions of formation content digests, that is

$$C_5 = \alpha_6(\alpha_5(\alpha_4(C_2, U_K), U_{CT}), U_D). \tag{3}$$

Relevant content is sent to a user and loaded into the database. The sampling distribution of the content is described like $C_6 = \alpha_7(C_5, U_{DS})$, where U_{DS} is the set of conditions of the sampling distribution of content.

Selective distribution of content list $C_7 = \max(C_6)$ depends on the level of demand for this content. Associative rule of formation content list consists of original content list C_6 and the content list which is selected from original and derived list C_7, in other words $C_6 \to C_7$. The formation of associative rule is a formation of content list ,that was formed through the merger of the original and derived lists. Associative rule of the appearance of the content from the original list along with the content from the list in the database is the following operator

$$P = \varphi(C_6, C_7) = \frac{\max(C_7 \cup C_6)}{\max(C_6)}, \tag{4}$$

where $d_s = \max(C_6)$ is the maximum of C_6 content set when $\forall c_{6i} \in C_6 \Rightarrow c_{6i} \le d_s$; $d = \max(C_7 \cup C_6)$ is the maximum $C = C_6 \cup C_7$ with $\forall c_i \in C = (C_6 \cup C_7) \Rightarrow c_i \le d$, then $d_s = \max(C_6) \Leftrightarrow d_s \in D_{C_6} \wedge \forall c_{faset} \in D_{C_6} : d_s \le c_{faset}$,

$$\begin{aligned} & d = \max(C_6 \cup C_7) \Leftrightarrow \\ & \Leftrightarrow d \in D_{C_6 \cup C_7} \wedge \forall c_{faset} \in D_{C_6 \cup C_7} : d \le c_{faset}, \end{aligned} \tag{5}$$

where $D_{C_6} = \{c_{faset} \in C_{faset} | \forall c_{6i} \in C_6 : c_6 \le c_{faset}\}$ is the set of maxima values for C_6 commercial Web-content at values of cardinalities set $\rho_{faset} = |C_{faset}|$, $\rho_6 = |C_6|$, $\rho_{faset} \ge \rho_6$; $D_{C_6 \cup C_7} = \{c_{faset} \in C_{faset} | \forall c_i \in (C_6 \cup C_7) : c_i \le c_{faset}\}$ is the set of maxima for $C_6 \cup C_7$ with $C_{faset} \ge (C_6 \cup C_7)$.

Indicators of profitability and growth of demand for commercial content are used to determine its relevance and calculate respectively as

$$I_p = \kappa(C_7, P) = P \frac{\max(C_6 \to C_7)}{\max(C_7)}, \tag{6}$$

$$I_g = \chi(C_7, P) = \frac{\max(C_6 \to C_7) - \max(C_7)}{\max(C_6 \to C_7)[1 - P]}. \tag{7}$$

Derived list of rules determines the moderator. The list formation of which exceeds the minimum level is the most used content list. Even if this condition is formed by a large number of lists of frequently used content in accordance with requests. The result is limited when the sample by operators

$$I_r = \psi(C_6, C_7, P) = \min(\forall C_6' \subset C_6, P - P'), \tag{8}$$

$$I_r = \min(\forall C_6' \subset C_6, \varphi(C_6, C_7) - \varphi(C_6', C_7)). \tag{9}$$

2. Method of management of commercial content. It is a support measures for

determining parameters of commercial Web-content (actuality, completeness, relevance, authenticity, reliability) in accordance with certain requirements by the set of criteria for the management of commercial Web-content.

Classification of management processes of commercial Web-content.

1. The management of commercial Web-content to generate pages on demand of the user of the ECCS is field as

$$\beta_Q = \langle C, Q, H, U, T, Z, \beta_1, \beta_2, \beta_3, \beta_4 \rangle . \tag{10}$$

The stage of editing and modification of Web-content is served by the operator $c_j(t_{r+1}) = \beta_1(c_j, t_r, h_k, u_l)$ with $c_j(t_{r+1}) \in C$. The development stage of pages set is described by operator $Z(t_r) = \beta_4(q_i, C, \beta_3(\beta_2(C), t_r))$, where

$$z_i = \left\{ \bigcup_{j=1}^{m} c_j(q_i, t_r) \left| \begin{array}{l} \forall c_j \in C_q, c_j \notin C_{\bar{q}}, C_q = \beta_3(\beta_2(C_q)), \\ \exists q_i \in Q_c, \exists h_k \in H_c, h_k \notin H_{\bar{c}}, \\ C = C_q \cup C_{\bar{q}}, Q_c \subset Q, H = H_c \cup H_{\bar{c}}, \\ k = \overline{1, n_H}, i = \overline{1, n}, r = \overline{1, w} \end{array} \right. \right\} . \tag{11}$$

The block weight is defined as the sum of the coefficients of the scales of commercial Web-content:

$$\omega = \|C\| = \beta_2(C, \omega_1, \omega_2, \omega_3, \omega_4, \omega_5), \tag{12}$$

where $\omega_1(c_j)$ is the coefficient of the location of the block in the commercial content, $\omega_2(c_j)$ is the coefficient of keywords in the block, $\omega_3(c_j)$ is the coefficient of statistical significance of terms, $\omega_4(c_j)$ is the coefficient of the availability of additional terms, $\omega_5(c_j)$ is the coefficient of the volume of terms from the user request.

2. Management of content with pages generation during editing Web- resource by moderator is presented as

$$\beta_E = \langle C, H, T, Z, \beta_1, \beta_2, \beta_3 \rangle . \tag{13}$$

Stage of pages set forming is described by operator

$$Z(t_r) = \beta_3(C, H, t_r, \beta_1, \beta_2) . \tag{14}$$

3. Mixed type of content management is presented as

$$\beta_M = \langle C, Q, H, T, Z, W, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \rangle , \tag{15}$$

where W is the set of cached content, β_5 is the operator of formation of the set of

cached Web-content or information blocks when $W = \beta_5(C, \beta_3(\beta_2(\beta_1(C, t_r, H, U)), t_{r+1})$ or $W = \beta_5(Z, \beta_3(\beta_2(\beta_1(C, t_r, H, U)), t_{r+1})$, where

$$w_l = \left\{ \bigcup_{i=1}^{n} c_i \middle| \forall c_i \in C_Q, C_Q \subset C, C_Q = \beta_3(\beta_2(C)) \right\}, \quad (16)$$

$$w_l = \left\{ \bigcup_{j=1}^{m} z_j \middle| \begin{array}{l} z_j \in Z_c, \forall c_j \in C_z, \exists c_j \in Z_c, \\ \forall c_j \in z_j, C_z = \beta_3(\beta_2(C)), \\ C_z \subset C, Z_c \subseteq Z, i = \overline{1,n} \end{array} \right\}. \quad (17)$$

3. Method of commercial Web-content supporting is set of measures providing functioning of the S e-business system according determined requirements and their following modifications. A analysis result of the S e-business system functioning and C commercial Web-content support is formed set $Y = \{Y_P, Y_T, Y_C, Y_R\}$ under the conditions $V = \{V_P, V_T, V_C, V_R\}$, where $Y_P = Y_{Pc} \cup Y_{Pq}$ is a subset of the information portraits of Y_{Pc} content and Y_{Pq} users, Y_T is a subset of thematic storyline of content, Y_C is subset of content relationship tables, $Y_R = Y_{Rc} \cup Y_{Rm}$ is a subset of the rating content Y_{Rc} and moderators Y_{Rm}, $V_P = V_{Pc} \cup V_{Pq}$ is the conditions set of information portraits formation, V_T is a conditions set for thematic storyline identification, V_C is the conditions set of the content relationship construct tables, V_R is the parameters set of the content ratings calculation. The information portraits set of content is presented as $Y_{Pc} = \gamma_1(V_{Pc}, C, H, Q, T)$, and set of the users` portraits are given as $Y_{Pq} = \gamma_2(V_{Pq}, Q, H, Z, T)$. The thematic storyline set for the content is presented as $Y_T = Y_{TC} \cup Y_{TQ}$, where $Y_{TC} = \gamma_3(C, H, X, V_T, T)$ is condition set for content storyline identification in the of new commercial content, and $Y_{TQ} = \gamma_4(C, H, Q, V_T, T)$ is thematic storyline definition. The set of relationship content tables is presented as $Y_C = \gamma_5(C, V_c, T)$. The set of the commercial Web-content rating is presented as $Y_{Rc} = \gamma_6(C, Q, H, Y_C, V_{Rc}, T, \theta, \xi)$, and a set of a moderators ratings Y_{Pq} is presented as $Y_{Rm} = \gamma_7(C, Q, H, Y_C, V_{Rm}, T)$, where $V_R = V_{Rc} \cup V_{Rm}$ is the parameters set for the content ratings calculation, $\theta(Q^+, Q^0, Q^-, T, H)$ is the tonality criteria for commercial content, $\xi(Q, T)$ is operator of comments filtering definition.

The Y set of output statistical data is presented as

$$Y = \{Y_P, Y_T, Y_C, Y_R\} = \gamma(V_P, V_T, V_C, V_R, C, Q, H, Z, T).$$

5 Research results analysis

The results chapter is to develop methods of processing Web-resources in the ECCS that make it possible to form the requirements for the creation, management and support of the Web-content. In the fourth chapter based on analysis of basic tasks of the *S* e-business system developed instrumental means and information technology to build such systems. The functional scheme of e-commerce content of subsystems of Web-resources processing is designed. The functional elements of the systems is described in detail overall architecture, objectives and principles of realization relevant systems and discuss. And scheme of action the most significant mechanisms is presented. Software of creation, management and support of the content is designed. There are posted software of developed systems implementation with subsystems of Web-resources processing (Tables 1, 2) in e-business organization over the Online Newspaper (ON) and Online Journal (OJ). Fig. 1-2 present the work results of the developed systems in the form of graphs. So the all stages presence of the commercial Web-content lifecycle significantly increases the visits and unique users amount of Web-resources.

Table 1. The results of systems operation

№	Web-resource	Address	Type
1	Fotoghalereja-vysocjkykh	fotoghalereja-vysocjkykh.com	OJ
2	Vgolos	vgolos.com.ua	ON
3	Tatjana	tatjana.in.ua	OJ
4	Presstime	presstime.com.ua	ON
5	AutoChip	www.autochip.vn.ua	OJ
6	Kursyvalyu	kursyvalyut.com	ON
7	Good morning	dobryjranok.com	ON
8	Information for Businesses	goodmorningua.com	OJ
9	LvivSchoolNumber3	зсш3львів.in.ua	OJ
10	Victana	victana.lviv.ua	OJ

Table 2. The results of systems operation in the time period from 10.2010 to 03.2014

Characteristic	1	2	3	4	5	6	7	8	9	10
% Content formation	30	100	10	40	20	90	70	60	0	50
% Content management	90	100	50	80	30	40	60	20	0	70
% Supporting content	30	100	10	40	20	50	80	60	0	70
% Uniqueness content	100	80	100	70	30	20	50	40	100	60
Visiting	4865	5997052	1381	3654456	9606	20132	8724	25	7	3138
Average duration of visit	4:41	2:14	3:56	2:04	1:51	1:02	2:27	8:12	0:46	4:15
% Denial Rate	56,14	71,90	53,2	83,08	55,67	82,9	68,15	48	97	32,9
GoalConversion Rate	7,83	0	0	0	12,51	0	0	0	0	0
Unique Visitors	3215	2501402	728	1501202	7105	16586	4996	7	5	1345
Page-views	22071	11588861	5464	769923	24908	31982	18892	81	12	18132
Pages / Visit	4,54	1,93	3,96	1,67	2,59	1,59	2,17	3,2	1,67	5,78
% New Visits	65,45	41,68	52,6	39,88	73,88	82,4	57,23	28	97,3	42,9

Characteristic	1	2	3	4	5	6	7	8	9	10
% Returning Visit	34,55	58,32	47,43	60,12	26,12	17,61	42,77	72	2,68	57,1
Source-not set	91,14	52,48	62,5	42,46	88,13	58,22	26,98	23	6,72	31,9
% Organic search	4,67	36,10	22,2	31,22	8,91	34,83	24,34	0	0,13	26,0
% Direct Traffic	2,14	11,20	7,53	26,12	2,35	5,34	48,73	77	93,12	27,9
% other sites	1,99	0,20	3,19	0	0,58	1,55	0,04	0	0,02	6,37
% social Network	0,06	0,02	4,56	0	0,02	0,06	0,01	0	0,01	7,81

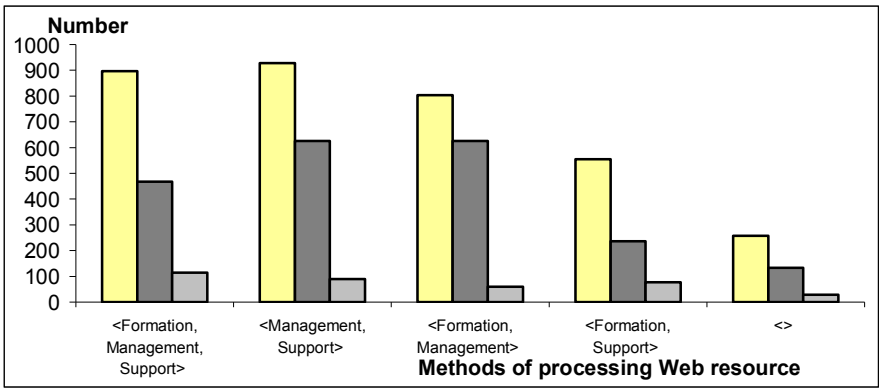


Fig. 1. Statistical analysis of the "Fotoghalereja-vysocjkykh" functioning.Value of fotoghalereja-vysocjkykh.com visiting and commercial content implementation according to the methods application (1 – Number of all visits of Web-resource, 2 – Number of visits of permanent users, 3 – Number of implementing content for permanent users)

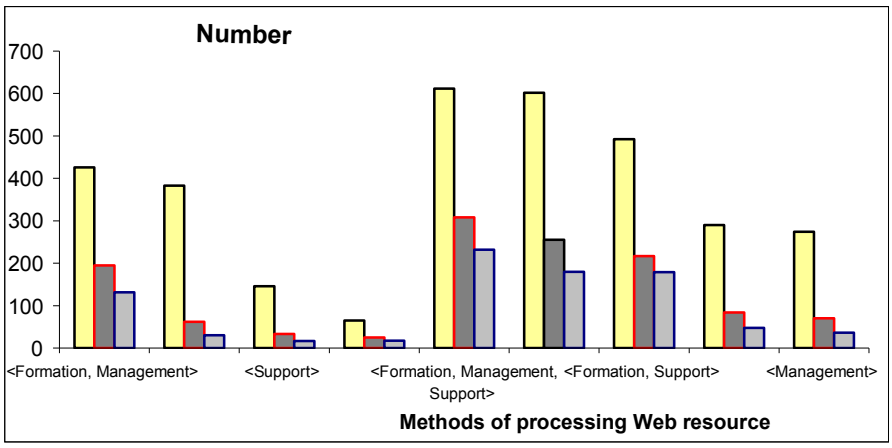


Fig. 2. Statistical analysis of functioning «Victana». Value of victana.lviv.ua visiting and commercial content implementation according to the methods application (1 – Web-resource visiting of the target audience, 2 – Web-resource visiting by permanent users, 3 – Commercial Web-content implementation of permanent users)

Service keeping statistics visits Web-resource allows us to estimate the increase in sales of commercial content which depending directly proportional increase in the

number of visits an information resource, the number of permanent users, the prospects of marketing events (Fig. 3).

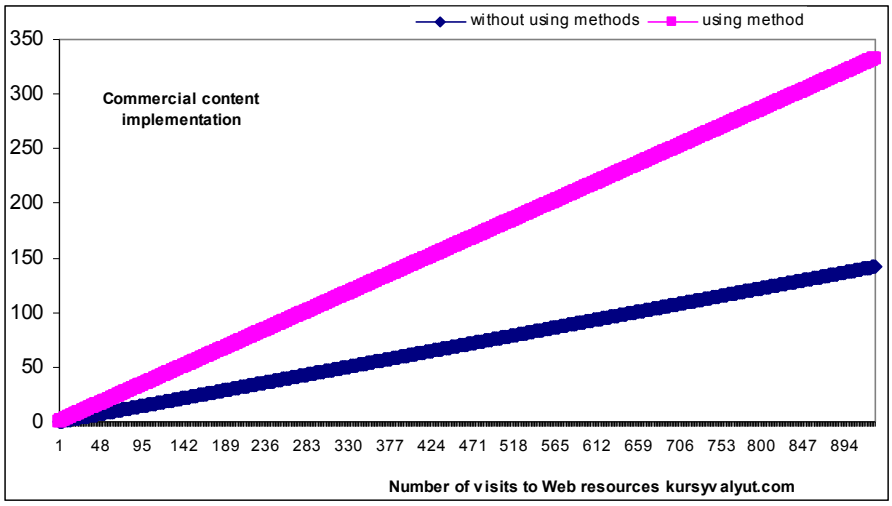


Fig. 3. Regressive analysis of increased sales of commercial content

The subsystems presence of Web-content creation, management and support in ECCS increases sales volume of content to the permanent user at 9%, active involvement of unique visitors, prospective users and expand the limits of the target and regional audience by 11%, viewed pages by 11%, visiting time and Web-resources by 9%. The results are the development of ECCS in the form of online newspaper and online magazine with the subsystems of formation, management and support of content.

6 Conclusion

The paper is solved the actual scientific problem of methods and means research and development for commercial Web-content processing in e-business systems. We obtain the following main scientific and practical results:

1. Investigated and improved classification of ECCS based on the analysis and evaluation of such systems, allowing us to determine, detail and justify their choice of functionality for the design life cycle of commercial Web-content.
2. For the first time developed a method for the formation of commercial content based on the improvement of its life to determine the commercial content flows control requirements, allowed us to automate the collection of data from different sources, identifying duplication and formatting commercial content, identifying key words and forming digests, selective dissemination of commercial content to improve its life cycle and requirements flows control commercial content.

3. Was improved method of control commercial content on the basis of its formation and analysis of the system to determine the values of control options to Web-content as relevant, aging, completeness, accuracy, relevance, authenticity, reliability.
4. For the first time developed a method for support commercial content based on statistical analysis of the functioning of e-commerce content to change the values of control options and requirements of forming commercial content that helped increase sales of Web-content in permanent users by 9%
5. Improved the structures of ECCS by analyzing the processing of Web-resources, different from the existing availability sub-formation, management and support of commercial content and prepare recommendations for the design of typical systems.
6. It is developed and implemented a software application of Web-content formation, management and support to achieve the effect of work at the level of the owner (profitability increasing, user interest growing) and user (clarity, the interface simplification standardization, unification, choice expanding), which made it possible increase the involvement of potential users and expand the boundaries of the target audience by 11%.

References

1. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2016).
2. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
3. CM Lifecycle Poster, Content Management Professionals, Access:http://www.cmprosold.org/resources/poster/ (2010)
4. Content Management Interoperability Services, Version V. 0.5, Hopkinton: EMC (2008)
5. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
6. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
7. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
8. Hackos, J.: Content Management for Dynamic Web Delivery. Hoboken: Wiley (2002)
9. Halverson, K.: Content Strategy for the Web. Reading: New Riders Press (2009)
10. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 47–55 (2017)
11. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)
12. McGovern, G., Norton, R.: Content Critical. Upper Saddle River: FT Press (2001)
13. McKeever, S.: Understanding Web content management systems: evolution, lifecycle and market. In: Industrial Management & Data Systems (MCB UP), 103(9), 686–692 (2003)

14. Nakano, R.: Web content management: a collaborative approach. Boston: Addison Wesley Prof. (2002)
15. Papka, R.: On-line News Event Detection, Clustering, and Tracking: thesis for the degree doctor of philosophy. Amherst: Massachusetts University (1999)
16. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
17. Rockley, A.: Managing Enterprise Content: A Unified Content Strategy. New Riders Press (2002)
18. Stone, W. Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! In: IEEE Antennas and Propagation, 45(4) 47-49 (2003)
19. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
20. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
21. The Content Management Possibilities Poster. http://metatorial.com/pagea.asp?id=poster.
22. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
23. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. . In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 100–109 (2017)
24. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)
25. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
26. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
27. Hamon, T.: Biomedical text mining. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf (2017)
28. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)
29. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
30. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
31. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)

32. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
33. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf. (2017)
34. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
35. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)
36. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
37. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 133–134 (2017)
38. Kirkin, S., Melnyk, K.: Intelligent data processing in creating targeted advertising. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 131–132 (2017)
39. Hordienko, H., Ilchenko, M.: Development and computerization of an English term system in the fields of drilling and drilling rigs. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 129–130 (2017)
40. Gorbachov, V., Cherednichenko, O.: Improving communication in enterprise solutions: challenges and opportunities. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 127–128 (2017)
41. Didusov, V., Kochueva, Z.: Statistical methods usage of descriptive statistics in corpus linguistic. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 125–126 (2017)
42. Verbinenko, Yu.: Discursive units in scientific texts. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 120–123 (2017)
43. Kotov, M.: NLP resources for a rare language morphological analyzer: danish case. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 31–36 (2017)
44. Kuprianov, Ye.: Semantic state superpositions and their treatment in virtual lexicographic laboratory for spanish language dictionary. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 37–46 (2017)
45. Titova, V., Gnatchuk, I.: Evaluation ofa formalized model for classification of emergency situations. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 110–119 (2017)

The Typhlocomments Rules for Audiodescription System of the Video Content Formation for People with Visual Impairments

Andriy Demchuk[1][0000-0001-5710-9347] and Olga Lozynska[2] [0000-0002-5079-0544]

[1]Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine, Andrii.B.Demchuk@lpnu.ua

[2]Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine, Olha.V.Lozynska@lpnu.ua

Abstract. This paper introduces the rules of constructing typhlocommentars for audiodescription system of the video content formation for people with visual impairments. This system can be considered as an intermediary between the available video content from the one side and the visually impaired user from the other. The process of forming the rules of typhlocomments is discussed. This rules are formed in the form of a proposal function, the truth of which is verified on the facts and the rules obtained as a result of psychological research. The use case diagram that describes the functional appointment of the system is suggested. The resulting use case diagram contains seven cases for use and two actors, among which the inclusion and extension relations are set. Tasks that require further research are defined.

Keywords: Video content, People with visual impairments, Typhlocomments, Typhlocommentar, Audiodescription.

1 Introduction

The development of human language and writing provided the undoubted help for visually impaired people. Partly, information holes can be filled in with a virtual description, oral or written language.

It can be assumed that the typhlocomments arose immediately after the appearance of the human language and at the same time with the appearance of the first person with visual impairment. There were three compelling reasons for this:

1. The presence of the visually impaired person, for whom there is a need to pass the information about the world by linguistic means.
2. The possibility of using the language that it enables the verbal method to describe a living entity, object, space or action in such a way that it can be understood without "seeing".

3. The presence of a sighted person who possesses the abilities to describe a verbal description for visually impaired persons (this person must be specially trained for this work).

Due to blindness, such persons have a problem with obtaining visual information and satisfaction of their educational-cognitive, cultural-aesthetic, integration-communicative and other socio-personal needs.

The typhlocomments is the targeted information specially prepared for people with visual impairment for replacement (or addition) of visual information perceived by a sighted person and which is unavailable (or inaccessible) to such persons due to loss of vision. Nevertheless, there are many known scientific articles on this issue. However, the problem of video content adaptation for persons with visual impairments has not been resolved yet.

The paper discusses an actual scientific and practical task of developing the typhlocomments rules for audiodescription system of the video content formation for people with visual impairments.

2 Related Work

Scientific research in the field of complementing video content with typhlocommentars to provide access to the video content for persons with visual impairments, began in the 80s of the last century and intensively evolve.

The basic theoretical foundations of the audiodescription are given in [1, 2]. S. Vanshin [3] introduced the notion of typhlocomments on post-soviet space, which is close to the concept of audiodescriptions in the rest of the world. In [4-6] methods of the adapted video content construction and its possible ways of development are given.

The scientists [7, 8] worked on information accessibility for visually impaired people in video format, but adaptation of the video content (finding places for insertion of typhlocommentars) has always been performed with the person help.

Development of science and using multimedia technologies have led to the creation of computer systems that provide access for people with visual impairments to information. Therefore, this situation can be overcome by developing methods and means of the video adaptation for visually impaired people.

3 Main Part

The audiodescription system can be considered as an intermediary between the available video content from the one side and the visually impaired user from the other.

There are three participants in forming of the video content for visually impaired people: the video content V (its owners), the visually impaired people K (clients) and the system for developing such content with the use of the audiodescription. For each

of representatives of these sets k, $(k \in K)$ and v, $(v \in V)$, their relations will be formed as suggestions and wishes for each of parties. Consider the main features of these objects.

The initial video content (without audiodescription) V has its own specifics, namely, it consists of the following elements:

- the total video content time;
- the non- dialogue time, where you can insert the audiodescription;
- the difficulty of describing events.

The total video content time clearly recognized by video content developers and contain information about the begin and the end of video and the soundtracks. It is described by the set of facts – logically true expressions that can be presented, for example, in the following form:

$Begin\_of\_sound \quad (v, 200\,s)$, $End\_of\_movie\,(v, 1\,hour\,20\,min)$.

The non-dialogue time is the time where you can insert the typhlocommentars. It determines the number of the parameters and the relates to the main possibilities of the developed system: the *S*-intervals for inserting typhlocommentars, the time of the typhlocommentars, the difficulty of the plot, the speed of reading. They are given by the set of pronouns in the form of predicates and each of them characterizes some property of the video content. Such relations are described with the use of the predicate constants:

$$Non-dialogue\ time(v, t <= 20s),$$

$$Time\ interval\ for\ audiodescription(v, 22\min, 24\min).$$

In round brackets, there are no arguments of the corresponding propositional functions, because their number and meaning depend on the particular predicate.

The difficulty describing events (plot). The typhlocommentar is formed during the revision of the video content. The peculiarity of forming the plot description is that it is mostly difficult to formalize the video content, but rather, it is necessary to describe it in some way in general. Such a description may be limited by statements of approximately this type:

"*All mass scenes need to be described with audiodescription*".

Uncertainty in user terminology and wishes requires the creation of the intellectual subsystem of the plot descriptions based on ontologies [9-11] and the participation of the expert to formulate and refine based on user suggestions rules [12-29]. The formulation of such rules is an important moment, since further approaches to the audiodescription should take into account the wishes of users (visually impaired persons) [30-39].

The suggestions and wishes of users *k* consist of:

- the objectivity of the plot description;
- no overlapped the audiodescription on dialogs.

The task of developing a high-quality video content for the user k based on the V following in satisfaction the above rules that give the true value of the logical function:

$$Video\ content\ OK(v,k).$$

During the practical implementation of the video content for visually impaired people, the system of audiodescription puts and takes into account in the further activity its experience in the form of knowledge about the user's psychology based on the perception of the sounding movies. Taking into account such knowledge may consist in the formation of some set of the rules. Such rules are formed in the form of a proposal function $D(v,k)$, the truth of which is verified on the facts and the rules obtained as the result of psychological research.

Finally, the success of the resulting video content for visually impaired people is to prove the truth of the goal function:

$$W(v,k) = Video\ content\ OK(v,k) \wedge D(v,k).$$

The conceptual model of the complex system, using use case diagram shown on Fig. 1.

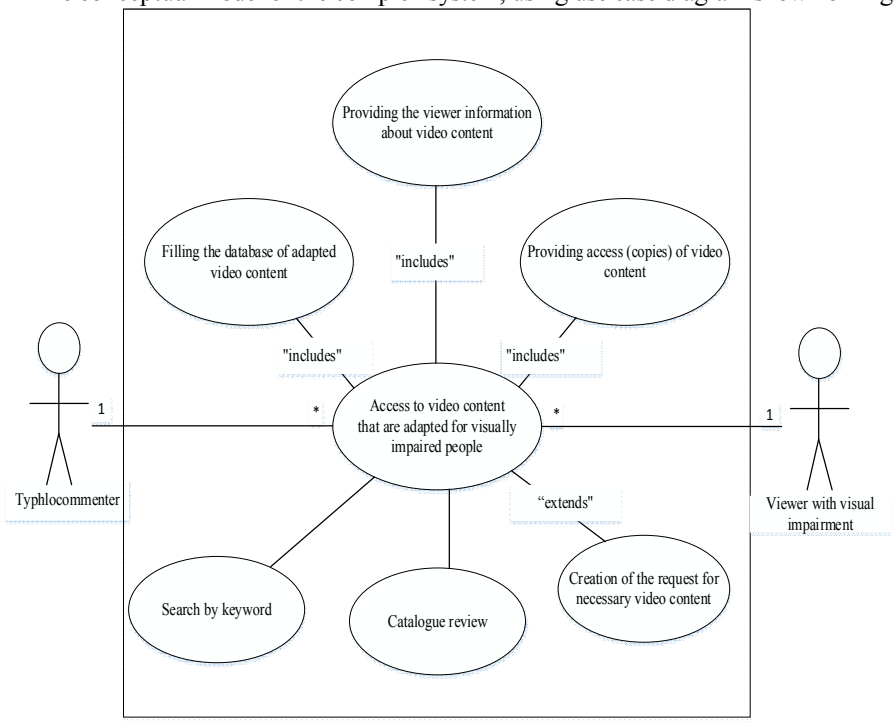


Fig. 1. The use case diagram of the audiodescription system

The use case diagram describes the functional appointment of the system. The goal of this diagram is that the projected system is presented in the form of the set of actors that interact with the system with using of so-called variants of use. In this case, the actor is called any entity that interacts with the system from the outside. This can be a

person, a technical device, a program, or another system that is a source of action on the simulated system. This action is defined by the system developer. In turn, the use case diagram is to describe the services that the system provides to the actor. In other words, each use case determines some set of actions carried out by the system during dialogue with the actor.

4 Results

The resulting use case diagram contains seven cases for use and two actors, among which the inclusion and extension relations are set. Access to the video content for viewers with visual impairments is realized through three cases: the keyword search, the catalogue review or the creation of the request for the necessary video content that must be adapted. Next, the viewer chooses the necessary action: gets information about the video content, gets access to it (opens / downloads it).

The audiodescription system is the part of program and algorithmic complex of video content adaptation for people with visual impairments. The approbation of program complex implementation results is made. The use of the rules of typhlocomments for the video content plot description allows to increase up to 30 % the perception of video content by people with visual impairments.

5 Conclusion

Loss of vision becomes a perceptible information barrier for people with visual impairments when visiting the museums and the exhibition halls, the theaters, the cinemas, the sports and the other cultural events, and makes it impossible to completely perceive the beauty of the works of art, the architecture, the literature, which is cultural and the historical heritage of mankind. The audiodescription system of the video content formation allows to partially solve this problem for person with visual impairments.

The process of forming the rules of typhlocomments is proposed. This rules are formed in the form of the proposal function, the truth of which is verified on the facts and the rules obtained as the result of psychological research. The use case diagram that describes the functional appointment of the system is suggested.
Further research can be focused on improving the audiodescription system.

References

1. Cintas, J. D., Remael, A.: Audiovisual translation: Subtitling. Across Languages and Cultures, 9(2), 291-299 (2008).
2. Siegel, B.: The world of the autistic child: Understanding and treating autistic spectrum disorders. New York, NY, US: Oxford University Press, (1996).
3. Vanshin, S. N., Vanshina, O. P.: The verbal description for the blind. M.: Logosvos,62 (2011).

4. Szarkowska, A.: Auteur Description: from the director's creative vision to audiodescription. Journal of Visual Impairment & Blindness, Vol. 7, 383-387(2013).
5. Frederick, W.,Moss, Jr.: Quality of Experience in Mainstreaming and Full Inclusion of Blind and Visually Impaired High School Instrumental Music Students. Michigan: University of Michigan, 258 (2009).
6. Audio Description for people with vision Loss. A Guide for Performing Arts Settings. John F. Kennedy Center for the Performing Arts, (2013). mode: http://www.kennedy-center.org/accessibility/education/lead/2013KC_AudioDescripGuide.pdf.
7. Igune, G. W.: Inclusion of Blind Children In Primary Schools: A case study of teachers' opinions in Moroto district – Uganda. Norway: University of Oslo, 108 (2009).
8. Inclusion: Creating an Inclusive Environment. A Handbook for the Inclusion of People with Disabilities in Nationaland Community Service Programs. – Access mode: http://www.service andinclusion.org/handbook/inclusion.pdf.
9. Demchuk, A. B.: Videocontent for the blind: the method of typhlocomments. Radio Electronics, Computer Science, Control: Scientific Journal Zaporizhzhya National Technical University,1 (30), 146-149 (2014).
10. Davydov, M., Lozynska, O.: Spoken and sign language processing using grammatically augmented ontology. Applied Computer Science 11(2), 29-42 (2015).
11. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Germany: LAP LAMBERT Academic Publishing (2016).
12. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
13. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
14. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
15. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
16. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
17. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
18. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
19. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
20. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 100–109 (2017)
21. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)
22. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)

23. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
24. Hamon, T.: Biomedical text mining. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf (2017)
25. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 47–55 (2017)
26. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)
27. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)
28. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
29. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
30. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)
31. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
32. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf. (2017)
33. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
34. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)
35. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
36. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 133–134 (2017)
37. Kirkin, S., Melnyk, K.: Intelligent data processing in creating targeted advertising. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 131–132 (2017)
38. Hordienko, H., Ilchenko, M.: Development and computerization of an English term system in the fields of drilling and drilling rigs. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 129–130 (2017)
39. Gorbachov, V., Cherednichenko, O.: Improving communication in enterprise solutions: challenges and opportunities. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 127–128 (2017)

Sentiment Analysis of the US and Ukrainian Presidential Speeches

Marianna Dilai[1], Yuliya Onukevych[2], Iryna Dilay[3]

[1-2]Applied Linguistics Department, Lviv Polytechnic National University
Lviv, Ukraine
[3]English Philology Department, Ivan Franko National University of Lviv
Lviv, Ukraine

mariannadilai@gmail.com[1],julia.bobko@gmail.com[2],
irynadilay@gmail.com[3]

Abstract. The paper presents the results of the sentiment analysis of the US and Ukrainian presidential speeches. By means of SentiStrength and UAM Corpus Tool programs, we attempt to extract opinions and sentiments in the speeches of Donald Trump and Petro Poroshenko. The main contribution of this study is the adaptation of the SentiStrength program to the Ukrainian language by compiling a political domain glossary of the Ukrainian emotion-bearing words. Furthermore, we compare lexical means of sentiment expression in the analyzed texts.

Keywords: sentiment analysis, Ukrainian sentiment lexicon, manual annotation, presidential speeches.

1 Introduction

The attention given to the research on the role of emotions in politics has grown dramatically over the past decades [1], [2], [3]. Numerous findings show that emotions in politics have distinct impact on public attitude and behavior, the direct consequences of which are drawing support and voting during the elections. Political psychology focuses on how emotions shape the evaluation of socio-political objects and affect information perception and processing [4]. Emotions are understood as appraisal-caused solutions, i.e. individuals appraise the significance of a political situation and act accordingly [5]. The application of sophisticated new technologies is likely to reveal the nuances of emotional appeals in politics. This study is conducted in line with modern trends in applied and computational linguistics, employing the sentiment analysis programs to extract and compare the usage of emotional appeals during the presidential campaigns in the USA (2016) and Ukraine (2014). Sentiment analysis is the study of the subjectivity (neutral vs. emotionally loaded) and polarity (positive vs. negative) of a text [6]. The research material is 50 speeches of the

President of Ukraine P. Poroshenko (52,248 words) and 50 speeches of the US President D. Trump (76,330 words).

In this paper we will combine automatic and semi-automatic sentiment analysis techniques. The first experiment is carried out by means of a SentiStrength algorithm[1] to automatically identify the polarity of sentiment in presidential speeches and detect the strength of a sentiment expressed. We will adapt the SentiStrength program to the Ukrainian language by compiling a political domain glossary of the Ukrainian emotion-bearing words. To process statistics SPSS (Statistical Package for the Social Sciences) will be used. In addition, taking into account the fact that emotional loading of words is context dependent, attitudinal meanings will be extracted within utterances by building a manually annotated dataset using the UAM Corpus tool 3.3h[2]. Finally, we will compare lexical means of sentiment expression in the texts of political speeches of the presidents of Ukraine and the USA. The algorithms of computational treatment of opinions, sentiment and subjectivity can help understand the fundamental role of emotion in the political discourse and identify affective utterances potentially associated with people's opinions and behaviors.

2 Previous research

Due to a wide set of applications, particularly in business and social sphere, sentiment analysis is one of the most in-demand issues of natural language processing. Sentiment analysis software can process thousands of text documents for subjective content and attitudes expressed toward a particular entity in seconds. Sentiment analysis algorithms aim at extracting opinions expressed in a text, determining the sentiments they convey, and classifying their polarity (positive, negative or neutral) as well as strength or intensity. Since early 2000 multiple techniques for sentiment analysis have been proposed, including lexicon-based approaches (e.g., General Inquirer, WordNet Affect, QWordNet or SentiWordNet) and supervised machine learning methods (e.g., Naive Bayes, MaxEnt, Support Vector Machine). Sentiment analysis can be applied at the discourse level, which presupposes that each document expresses opinions on a single entity [7]. The sentence-level sentiment analysis determines whether the sentence implies positive or negative opinions. The object-oriented sentiment analysis reveals sentiment towards a specific entity mentioned in the text [8]. The aspect-based sentiment analysis focuses on opinions relative to specific properties (or aspects) of an entity [9], [10].

However, the abovementioned applications and enhancements on SA algorithms are mostly limited to the analysis of English texts. Not much research has been done in this field in Ukraine. Some contributions to the development of sentiment analysis of Ukrainian texts were made by M. Lobur, A. Romaniuk, M. Romanyshyn [11].

---

[1] SentiStrength, http://sentistrength.wlv.ac.uk.

[2] UAM Corpus tool, http://www.corpustool.com.

3 Methodology

In course of this research, in order to reveal emotional loading of presidential speeches, we are to carry out two sentiment analysis experiments. The first one is performed applying a SentiStrength program. It is a lexicon-based sentiment analysis of the texts of presidential speeches which deals with text segments (sentences) and, besides sentiment-marked words, takes into account some text features, which affect the emotive load of the sentence. The second experiment lies in manual segment annotation of the texts using UAM Corpus Tool.

Experiment 1. The SentiStrength program determines positive and negative polarity of words and their strength on a scale ranging from -5 to 5 [12]. Based on the research from psychology, which has revealed that people process positive and negative sentiment in parallel, SentiStrength reports two sentiment strengths: -1 (not negative) to -5 (extremely negative), 1 (not positive) to 5 (extremely positive),. The SentiStrength algorithm is based on the information given in the files, including:

- EmotionLookUpTable.txt, which contains a list of emotion-bearing words, each one with the word then a tab, then an integer 1 to 5 or -1 to -5. Strengths of +1 and -1 have no effect on the program.
- NegatingWordList.txt, which reverses the polarity of subsequent words, e.g. not happy is negative.
- BoosterWordList.txt, which increases sentiment intensity, e.g. very happy is more positive than happy.
- IdiomLookupTable.txt, which overrides the sentiment strength of the individual words in the phrase

The SentiStrength analysis output is a copy of the text file with overall comment assigned with the most positive of its sentence emotions and the most negative of its sentence emotions. Figure 2 shows the output for D. Trump's speeches transformed into Excel.

	A	B	C	D	E	F	G	H	I	J	K
1		Translation	Positive	Negative	Sum	EmotionRationale					
2	Thank you, Mel	Thank you, Mela	3	-1	2	Thank[1] you[0] Melania[0] for[0] that[0] wonderful[2] introd					
3	Secretary Wilso	Secretary Wilso	3	-2	1	Secretary[0] Wilson[0] General[0] Goldfein[0] Chief[0] Wrigh					
4	Before we begi	Before we begir	3	-4	-1	Before[0] we[0] begin[0] I[0] want[0] to[0] say[0] that[0] our[					
5	America and ou	America and our	4	-4	0	America[0] and[0] our[0] allies[0] will[0] never[0] be[0] intim					
6	After seeing yo	After seeing you	2	-3	-1	After[0] seeing[0] your[0] capabilities[0] and[0] commitment					
7	Our thoughts al	Our thoughts als	2	-1	1	Our[0] thoughts[0] also[0] remain[0] with[0] those[0] recover					
8	And we know w	And we know w	3	-2	1	And[0] we[0] know[0] we[0] can[0] always[0] count[0] on[0] t					
9	I am truly thrille	I am truly thrille	4	-2	2	I[0] am[0] truly[1] thrilled[3] to[0] join[0] you[0] today[0] at[0					
10	We're celebrati	We're celebratir	4	-1	3	We're[0] celebrating[1] 70[0] years[0] of[0] history[0] 70[0] y					
11	For seven deca	For seven decad	2	-3	-1	For[0] seven[0] decades[0] the[0] United[0] States[0] Air[0] F					

Fig. 1. The SentiStrength analysis of D. Trump's speeches

The SentiStrength program has been adapted to the Ukrainian language by replacing the English term list EmotionLookupTable.txt, NegatingWordList.txt, BoosterWordList.txt, and QuestionWords.txt with the Ukrainian ones.

The word list EmotionLookupTable.txt was compiled by translating the English word list and adding the most common sentiment-bearing words from the analyzed speeches. Each word is marked with a score that indicates the polarity and the intensity of the sentiment expressed -5, -4, -3, -2, 2, 3, 4, 5. (-1, 0, and 1 are not used). In order to get balanced results, 50 people of different age and gender were asked to classify the words ranking them taking into account the typical use of the word in political context. Then, each word was given an average score. The functional affixes of words which do not change the sentiment meaning wee truncated and replaced with a star *. After that the word list was incorporated in the program. The final list consists of 3,000 words.

We use the Java version that needs the utf8 option to read the input files. Figure 2 presents the results of P. Poroshenko's speeches in Excel.

	B	C	D	E	F
1	Translation	Positive	Negative	Sum	EmotionRationale
2	Шановні українці!	1	-1	0	Шановні[0] українці[0] [[Sentence=-1,1=word max, 1-
3	Сьогодні Україна відзначає	3	-2	1	Сьогодні[0] Україна[0] відзначає[0] День[0] пам'яті[
4	Висловлюю нашу глибоку п	3	-2	1	Висловлюю[0] нашу[0] глибоку[0] повагу[0] і[0] шан
5	Коли наші діди й прадіди в	1	-1	0	Коли[0] наші[0] діди[0] й[0] прадіди[0] в[0] 44[0] му
6	Прошу, погляньте, на цих по	3	-1	2	Прошу[0] погляньте[0] на[0] цих[0] портретах[0] обл
7	«Четверте покоління» – діти	3	-1	2	«Четверте[0] покоління»[0] –[0] діти[0] незалежної[
8	Цю виставку слід сприймати	3	-1	2	Цю[0] виставку[0] слід[0] сприймати[0] як[0] симво
9	На жаль, в Європі – знову ві	1	-2	-1	На[0] жаль[-1] в[0] Європі[0] –[0] знову[0] війна[-1]
10	Сьогодні зі знятими погона	4	-1	3	Сьогодні[0] зі[0] знятими[0] погонами[0] та[0] зама
11	Натомість для всього україн	3	-4	-1	Натомість[0] для[0] всього[0] українського[0] суспіл
12	Нещодавно до мене зверну	2	-1	1	Нещодавно[0] до[0] мене[0] звернулися[0] десятки
13	Сьогодні тут присутні предс	4	-1	3	Сьогодні[0] тут[0] присутні[0] представники[0] гром
14	Вважаю, кращої історичної	3	-2	1	Вважаю[0] кращої[0] історичної[0] локації[0] для[0]
15	На мою думку, такий підхід	3	-1	2	На[0] мою[0] думку[0] такий[0] підхід[0] утверджува
16	Слава ветеранам Червоної	2	-1	1	Слава[1] ветеранам[0] Червоної[0] армії[0] та[0] пар
17	Слава воякам Української п	2	-1	1	Слава[1] воякам[0] Української[0] повстанської[0] а
18	Слава всім, хто захищав Укр	2	-1	1	Слава[1] всім[0] хто[0] захищав[0] Україну[0] від[0]

Fig. 2. The SentiStrength analysis of P. Poroshenko's speeches

The results of the automatic sentiment analysis are compared with the manual and semi-automatic linguistic annotation of texts using UAM Corpus Tool.

Experiment 2. The UAM Corpus tool 3.0.enables segmented manual annotation using built-in Appraisal framework by Martin & White 2005, whose central part is Attitude system (see Figure 3). Affect, judgment, and appreciation are basic types of attitude. Affect expresses a person's internal emotional state. Judgment evaluates a person's social behavior in a context. Appreciation evaluates norms about how things, performances, and naturally occurring phenomena are valued, when this evaluation is expressed as being a property of the object. The attitude can be of positive or negative polarity. The attitude segments can be expressed implicitly or explicitly.

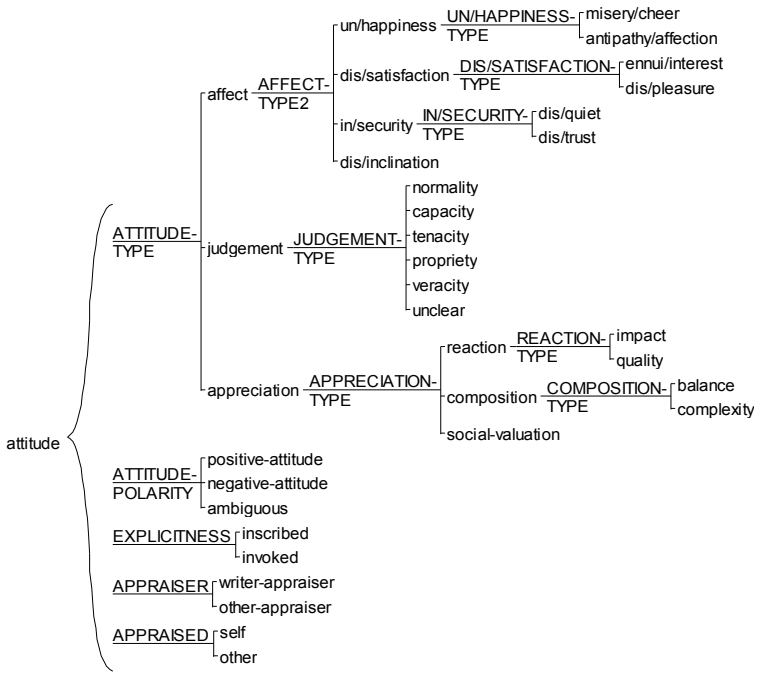


Fig. 3. Attitude Scheme by J. Martin and P. White

A manually annotated dataset allows taking into account the context determining emotional appeal in utterances. Figure 4 shows a fragment of the UAM annotation of P. Poroshenko's speeches.

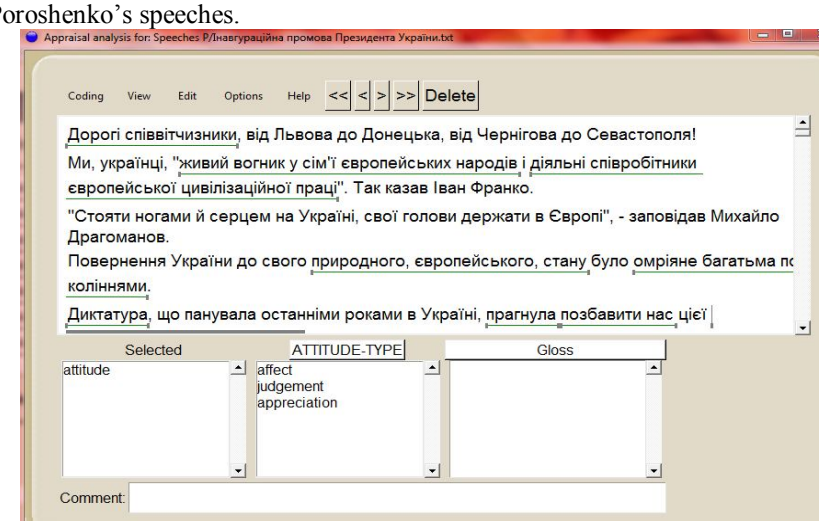


Fig. 4. UAM Corpus tool annotation of P. Poroshenko's speeches

Here are the examples of the annotation of some attitude-bearing segments in D. Trump's and P. Poroshenko's speeches:

(1) They have to do, finally, what's right for the American people. But probably we'll do it ourselves. Because today ***we won*** *51 to 50 and didn't get one Democrat vote. Think of that.*

We won**:** affect – happiness – (misery) cheer – positive attitude – invoked – writer appraiser – self;

(2) It looks like about 45,000 people. ***You set a record*** *today.* ***You set a record****. That's a great honor, believe me.*

You set a record: affect – (dis)satisfaction – (dis)pleasure – positive attitude – invoked – writer appraiser – self;

(3) ***Слава Україні!***

Слава Україні*:* affect – pleasure – (dis)satisfaction – positive attitude – invoked – writer appraiser – other

(4) Диктатура, що панувала останніми роками в Україні, ***прагнула*** *позбавити нас цієї перспективи – народ повстав.*

Прагнула: affect – (dis)inclination – positive attitude – inscribed – writer appraiser – other.

4 Results and discussion

4.1 The SentiStrength outcomes

It has been revealed that D. Trump's speeches are highly emotional (26,8 % of the text segments are neutral). The positive sentiment prevails over the negative one, making up 46.9% and 26.3% correspondingly (see Table 1, Figure 5).

Table 1. The SentiStrength results: D. Trump's speeches

		Frequency	Percentage %
Scale	-4	5	.4
	-3	40	2.9
	-2	95	6.8
	-1	226	16.2
	0	373	26.8
	1	357	25.7
	2	245	17.6
	3	46	3.3
	4	4	.3
	Total	1391	100.0

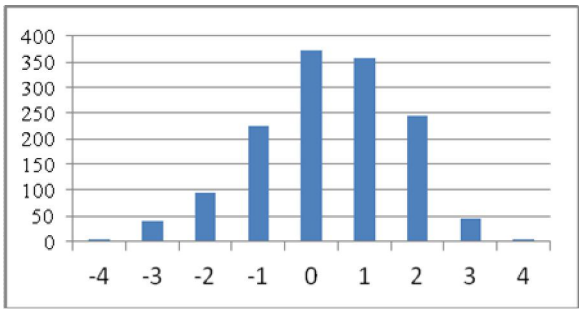


Fig. 5. The SentiStrength results: D. Trump's speeches

The most frequently used words by D. Trump are given in Table 2.

Table 2. The most frequently used words by D. Trump

Word	Index	Count
thank	1	343
great	2	313
like	1	171
tax	-1	167
good	1	110
applause	1	109
love	2	109
proud	1	60

The SentiStrength results for P. Poroshenko's speeches show that they are more neutral (54.7% of neutral text segments). The positive polarity considerably prevails the negative one, the program identified 37.9% of positive segments and 13.3% of negative segments (see Table 3, Figure 6).

Table 3. The SentiStrength results: P. Poroshenko's speeches

		Frequency	Percentage %
Scale			
	-4	3	.1
	-3	39	1.9
	-2	58	2.8
	-1	173	8.5
	0	1119	54.9
	1	432	21.2
	2	139	6.8
	3	61	3.0
	4	14	.7
	Total	2038	99.7

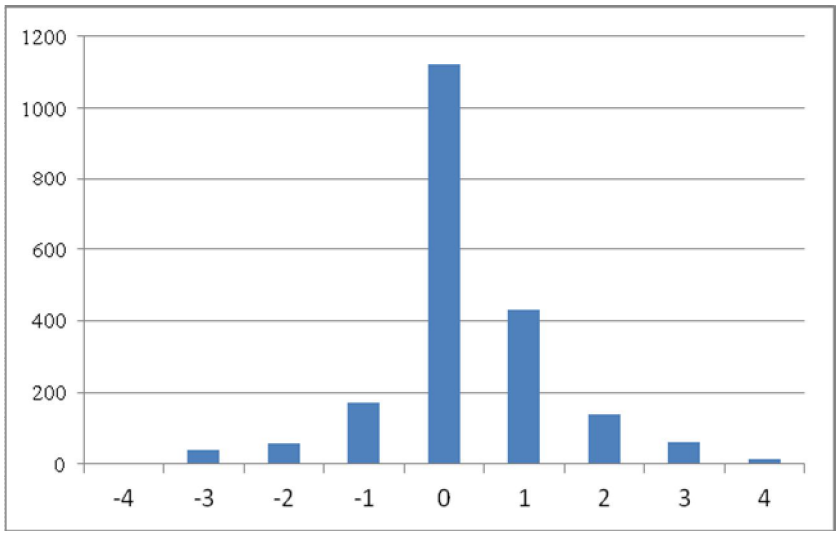


Fig. 6. The SentiStrength results: P. Poroshenko's speeches

The most frequently used words by P. Poroshenko are given in Table 4.

Table 4. The most frequently used words by P. Poroshenko

Word	Index	Count
проти	-1	62
слава	1	60
жаль	-1	35
свободи	1	35
свободу	1	34
впевнений	2	34
війни	-3	32
подякувати	1	30
героїв	2	30
важливо	1	29
друзі	1	26
мир	2	25

Both presidents mainly use fairly positive words with index 1. However, P. Poroshenko, unlike D. Trump, tends to use extremely positive words marked 4. The most common words in terms of intensity are shown in Table 5.

Table 5. The most common words in terms of intensity

Index	D. Trump's words	P. Poroshenko's words
-4	*devastation, devastating, excruciating,*	*вбивство, вбити*
-3	*violence, horrible, terrible*	*війни, ворог, біль*
-2	*fight, fighting, terrorists*	*жертв, геноциду, агресор*
-1	*tax, against, bad*	*проти, жаль, боротьба*
1	*thank, like, good*	*слава, свободи, свободу, подякувати*
2	*great, love, beautiful*	*впевнений, мир, героїв, цінності*
3	*terrific, thrilled, magnificent*	*добре, радий, честь*
4		*правда, задоволення, успіх*

4.2 The UAM Corpus Tool outcomes

The manual annotation made using UAM Corpus Tool 3.3h shows that both D. Trump's and P. Poroshenko's speeches most frequently express affect (with identical frequency of 46%) and have positive polarity (77.43% and 73.77% correspondingly). The results of UAM Corpus Tool analysis are shown in Figures 7, 8 and Table 6.

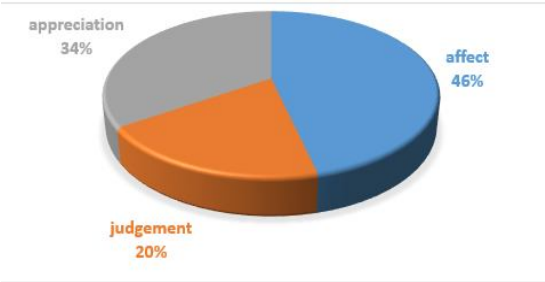


Fig. 7. The UAM Corpus Tool outcomes for D. Trump's speeches

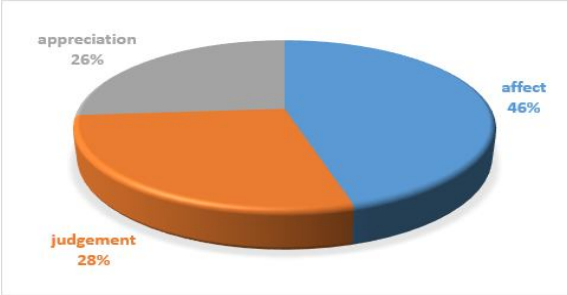


Fig. 8. The UAM Corpus Tool outcomes for P. Poroshenko's speeches

Table 6. The UAM Corpus Tool results for D. Trump's and P. Poroshenko's speeches

Feature	D. Trump's speeches	P. Poroshenko's speeches
Polarity	Positive – 77.43% Negative – 20.19% Ambiguous – 2.40%	Positive – 73.77% Negative – 25.01% Ambiguous – 1.22%
Explicitness	Inscribed – 80.98% Invoked – 19.02%	Inscribed – 91.38% Invoked – 8.62%
Appraiser	Writer-appraiser – 85.75% Other-appraiser – 14.25%	Writer-appraiser – 98.33% Other-appraiser – 1.67%
Appraised	Self – 33.33% Other – 66.67%	Self – 28.66% Other – 71.34%

The manual annotation shows that the presidential speeches of P. Poroshenko and D. Trump have many features in common. In particular, attitude often implies affect and is of a positive polarity. However, Donald Trump uses implicit attitude more often (irony, sarcasm, figurative meaning, etc.), whereas Petro Poroshenko often quotes Ukrainian proverbs or prominent people in various fields in his speeches.

4.3 Correlation between the two programs

SentiStrength and UAM Corpus Tool have revealed that positive sentiment prevails over the negative one in the speeches of both presidents. The difference in percentage rate can be explained by the fact that UAM Corpus Tool output does not measure the frequency of neutral segments. In addition, the SentiStrength analysis does not take into account the context in which a word is used and figurative language, including metaphorical expressions, proverbs and sayings, which abound in political texts and are highly emotional. Moreover, the SentiStrength program fails to differentiate between homonyms. In example (6), the word like is not emotionally loaded. Similarly, the word єдині in example (8) is neutral.

(5) From the earliest wooden biplanes, to the high-tech UAVs, to the awesome power and stunning beauty of the F-35, B-2, F-22, – and I saw a lot of them today – the F-15, the F-16, the F-18, I don't know which one I ***liked*** *the most.*

(6) Those days are going to be over very soon. Unfortunately, ***like*** *everything else, there are so many restrictions on doing what's right for our country, there's so much red tape, but don't worry, we're getting through it.*

(7) Ми маємо бути ***єдиними****. Навіть найміцніші мури впадуть, якщо всередині фортеці панує внутрішній розбрат.*

(8) Освітня реформа не ***єдина****, якій ми даємо старт цієї осені.*

5 Conclusions

As a result of the experiments carried out based on the application of the automatic and semi-automatic sentiment analysis, it has been revealed that the presidential speeches of D. Trump and P. Poroshenko are subjective and of positive polarity.

Manual segment annotation enabled taking into account the context, homonymy and figurative language, including metaphorical expressions, which are typical of presidential speeches. The outcomes have shown that positive attitudes expressing affect prevail in the speeches of both presidents. The main contribution of the study is adjusting the SentiStrength program to the automatic analysis of the Ukrainian political texts, by incorporating the Ukrainian political domain emotion-bearing lexicon. The research findings can serve as the basis for the further advancement of the sentiment analysis methodology of Ukrainian texts.

References

1. G.E. Marcus, Emotions in politics, Annual Review of Political Science, 3, 2000, 221-50.
2. N. A. Valentino, T. Brader, E. W. Groenendyk, K. Gregorowicz & W. L. Hutchings, Election night's alright for fighting: The role of emotions in political participation, Journal of Politics, 73, 2011, 156-170.
3. B. Albertson & S. K. Gadarian, Anxious Politics: Democratic Citizenship in a Threatening World, 2015, New York: Cambridge University Press,.
4. M. Lodge & C. Taber, Implicit affect for political candidates, parties, and issues: An experimental test of the hot cognition hypothesis, Political Psychology, 26, 2005, 455-482.
5. C. A. Smith & P.C. Ellsworth, Patterns of cognitive appraisal in emotion, Journal of Personality and Social Psychology, 48, 1985, 813-838.
6. B. Pang & L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, Vol. 2, No 1-2 , 2008, 1–135.
7. B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol. 10, 2002,. 79–86.
8. E. Amigo, A. Corujo, J. Gonzalo, E. Meij, M. de Rijke, Overview of RepLab 2012: Evaluating Online Reputation Management Systems, CLEF 2012 Evaluation Labs and Workshop Notebook Papers, Rome.
9. B. Liu, Sentiment analysis and opinion mining, 2012, Morgan & Claypool Publishers.
10. L. Zhang, B. Liu, Aspect and Entity Extraction for Opinion Mining, in Data Mining and Knowledge Discovery for Big Data, 2014, Springer, Berlin, Heidelberg, 1–40.
11. M. Lobur, A. Romaniuk, M. Romanyshyn, Defining an approach for deep sentiment analysis of reviews in Ukrainian, Visnyk Natsionalnogo Universytetu Lvivska Politehnika, No 747, Komputerni systemy proektuvannia, Teoria i praktyka, 2012, 124–130.
12. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, & A. Kappas, Sentiment strength detection in short informal text, Journal of the American Society for Information Science and Technology, 61(12), 2012, 2544–2558.

The Quantitative Research of Scientific Texts at the Symbolic Level

Ihor Kulchytskyi, Ulyana Shandruk

Lviv Polytechnic National University, Lviv, Ukraine

Abstract. The concept of text has been one of the important aspects of the scientific research since the beginning of the last century. It has been incorporated in the scientific paradigms of various fields of human knowledge such as psychology, philosophy, science of culture, linguistics, pedagogy. The researchers call this terms interdisciplinary and they claim it to be one of the key concepts in the field of culture of language, linguistics, and communicative linguistics. Text is a unit of a written language that is used by immense audience, therefore, the difficulties of communication are growing with the increase in the number of people who actually communicate. So, the concept of text in linguistics has always remained one of the least investigated. The paper aims to contribute into the research of text in general and scientific text in particular and reveal if any the connection between the usage of symbols and the genre the investigated text belongs to.

Keywords: Text, Corpus, Scientific Text, Fiction Text, Symbols, Entropy.

1 Introduction

According to the researchers [4, 11, 12, 15], scientific text is a way of representing scientific information. It is also the result of the scientific research that:

- always reflects some problem, proposes hypotheses, creates new knowledge; is characterized by expediency and rationality; is oriented on the achievement of research goals and tasks;
- has a rational nature, consists of judgments and inferences; is constructed according to the rules of the logic of science and formal logic;
- is characterized by widespread usage of the conceptual and categorical apparatus of science;
- is not based on the image; does not activate the perception of the reader, but focuses on the sphere of rational thinking;
- its purpose is not to make it believe, but to prove, justify, and argue the truth.
- the linguistic features of scientific texts are the following: [11]:
- strict regulation (compliance with the norms of the standard language);
- the widespread usage of abstract vocabulary;

- impersonality;
- due to the content it is presented in the form of monologue;
- completeness and completeness of statements;
- a close connection between parts of the statement, the logical sequence;
- the usage of concise, but highly informative attribute-name phrases;
- the usage of symbols;
- the primacy of the written language;
- the absence of subtext, etc.

Another feature of scientific text that is necessary to mention is that they, much more than literary texts, involve some degree of uncertainty and subjectivity. Objective and accurate, as it may seem from the first glance, scientific text mostly deal with some hypotheses and assumptions [1]. In their papers scientists have to conclude on the data they process and such concussions are often uncertain. Sometimes they are not sure about their results, the other time they just propose some result they cannot interpret although they have done much work and even no results is a result. This distinction has led to the fact that the language the scientist use also reflects the uncertainty and subjectivity [3, 2].

An overview of the evolvement and the current state of research of scientific texts is proposed in the papers of Vysotka [4] and Ivasenko [10]. At the same time, much less attention is paid to the linguistic statistical parameters of scientific text. Since the language reveals its objective properties both in quantitative and qualitative respects, it is impossible to explain the functioning of a number of language categories without the use of statistical methods. The investigation of the quantitative characteristics of the lexical filling of text can make a significant contribution to the analysis of scientific text and the quantitative stylistics of scientific texts, which determines the topicality of our research. The aim is to define quantitative characteristics of scientific and technical texts and compare them with the similar characteristics in literary texts.

As a real object, the language has a complex structure. Usually it is considered as a hierarchical six-level structure (graphemes, phonemes, morphemes, lexemes, syntaxemes, and text). The units of the previous level are building the units of the next level [14]. Basically, starting from the phonetic to textual levels, the research there has been conducted in the proper way whilst we believe the research of graphemes is yet to be improved. The paper presents the results of the investigation of symbols in the different scientific texts.

2 The organization of the research material

The material for studying the quantitative characteristics of scientific texts at the symbolic level is the texts from the scientific journal «Information systems and networks» published by the department of «Information systems and networks» of the Lviv Polytechnic National University. The scientific journal, which is in the list of professional publications, is aimed to reflect the main results of dissertation research.

From 1997 the series is issued by the scientists of the department of "Information systems and networks". The thematic focus of the scientific journal is basically on technical sciences namely computer science, computer technology and automation, systems and processes of management, computer systems and components, Information Technology, automation of control processes, medical and biological informatics and cybernetics, systems of the projecting automation, information security systems, project and program management, systems and means of artificial intelligence, computer science and cybernetics, mathematical modeling and computational methods, the mathware and software of computing machines and systems, system analysis and theory of optimal solutions, structural, applied and mathematical linguistics.

Articles are divided into 2 groups according to the following headings: "Information systems and networks", "Computer and mathematical linguistics", "Project and program management".

The research embraces 11 scientific journals, containing 391 articles for the period of 2009-2016 [2]. This selection is due to the fact that since 2009 this scientific journal has received a standard structural formatting and division according to subject headings.

To compare the quantitative characteristics of the scientific texts with the quantitative characteristics of the literary texts, the following texts of R. Ivanychuk were used: “The fire pillars” [3], “Water from the stone” [4], “Manuscript from the Ruska street” [4], “Haloxylon in the sands” [4], “The goddess pilgrimage”[5], and “The red wine” [6]. The texts of these novels were converted into the electronic form and normalized [Куль 1] for further research. The choice of these works of fiction was due to the presence of their normalized texts which significantly reduced the preparation of the materials for the research.

At the first stage, the scientific texts were converted into electronic forms, edited and normalized [13]. The editing process involved removing random artifacts that got into the text in the process of the text recognition, and also correcting wrongly recognized words. During the normalization process, the following actions were taken:

- all symbols were codded according to Unicode (utf8);
- the coding of punctuation marks, including dashes and hyphens, was done according Ukrainian spelling and editorial regulations;
- the signs of breaking words on to a new line were removed;
- empty paragraphs were removed as well;
- the spaces at the beginning and at end of the paragraph were cut;
- only one space was left between words.

As mentioned above the research corpus was based on the materials from the scientific journal «Information systems and networks» [5] published by the department of «Information systems and networks» of the Lviv Polytechnic National University from 2009-2016. All articles were divided into several groups such as

"Information systems, networks and technologies" (ISN), "Computational and mathematical linguistics" (CML), "Projects and program management" (PPM) and the novels by R. Ivanychuk (the texts by R. Ivanychuk or TRI) namely "The fire pillars" [6], "Water from the stone" [7], "Manuscript from the Ruska street" [7], "Haloxylon in the sands" [7], "The goddess pilgrimage"[8], and "The red wine" [9]. The word occurrences mean the general number of all grammatical forms of the word in the text. Equal grammatical forms were counted as different word occurrences. The number of grammatical forms is the number of different grammatical forms that occurred in the text. Each grammatical form was counted only one time. The number of symbols is the general number of characters of the extended Ukrainian alphabet that includes the symbols of traditional Ukrainian alphabet, space, hyphen and apostrophe. The general characteristics of the corpus are presented below:

Table 1. The general characteristics of the investigated corpus.

Thematic group	The quantity			
	Texts	The number of word in one of its lexical meanings	The set of all grammatical forms of the word	Symbols
Information systems, networks and technologies	235	729060	59686	5152579
Computational and mathematical linguistics	101	347381	42748	2456785
Projects and program management	55	167240	24944	1274349
The novels by R. Ivanychuk	6	484729	80626	2983097

3 The conducting of the research

For the statistical research at the level of symbols, the four arrays of character strings were formed from the scientific texts divided into the thematic groups and novels by R. Ivanychuk. All the arrays were created using the following rules:

- only the characters of the extended Ukrainian alphabet were left in the texts;
- all letters are converted into uppercase letters;
- only once space was left between words;
- one large text was formed by joining together all texts from the thematic groups leaving the space between them;
- the formed array was divided into paragraphs of equal length of 500 characters. This is due to the research algorithms. The number 500 was chosen introspectively.

4 The investigation of the frequency of the symbols

In each of the four text arrays, the absolute and relative frequency of each character and its rank were calculated. The results are presented in Table 2.

Table 2. The frequency distribution of the symbols of the investigated selections.

	ISN	CML	PPM	TRI		ISN	CML	PPM	TRI
А	7,00%	6,75%	6,77%	7,04%	О	8,15%	8,32%	8,20%	8,15%
Б	1,18%	1,13%	1,14%	1,51%	П	2,46%	2,37%	2,45%	2,48%
В	4,22%	4,39%	4,22%	5,16%	Р	4,40%	4,10%	4,48%	3,65%
Г	1,28%	1,23%	1,27%	1,51%	С	3,56%	3,75%	3,48%	3,64%
Ґ	0,01%	0,02%	0,02%	0,02%	Т	5,22%	5,24%	5,24%	3,94%
Д	2,88%	2,78%	2,80%	2,72%	У	2,75%	2,80%	2,81%	2,91%
Е	4,45%	4,32%	4,45%	3,77%	Ф	0,51%	0,58%	0,53%	0,11%
Є	0,64%	0,57%	0,60%	0,36%	Х	1,17%	1,02%	1,18%	0,95%
Ж	0,66%	0,68%	0,66%	0,75%	Ц	1,01%	0,90%	1,13%	0,61%
З	2,14%	1,97%	2,09%	2,01%	Ч	1,01%	1,13%	0,94%	1,12%
И	5,11%	5,24%	5,04%	5,66%	Ш	0,41%	0,39%	0,39%	0,75%
І	5,40%	5,28%	5,62%	4,53%	Щ	0,27%	0,26%	0,27%	0,45%
Ї	0,65%	0,66%	0,66%	0,49%	Ь	1,37%	1,38%	1,41%	1,42%
Й	0,79%	0,79%	0,84%	1,42%	Ю	0,70%	0,76%	0,72%	0,76%
К	2,95%	3,22%	3,03%	3,03%	Я	2,14%	2,05%	2,14%	2,03%
Л	2,61%	2,88%	2,63%	3,36%	‘	0,15%	0,14%	0,15%	0,10%
М	2,90%	2,98%	2,84%	2,65%	-	0,24%	0,26%	0,23%	0,07%
Н	7,51%	7,34%	7,57%	5,12%	sp	12,11%	12,32%	12,00%	15,78%

The analysis carried out during the research has enabled to highlight the groups of the symbols with high, medium and low frequencies.

The group of the most frequent symbols (the space (sp) is always the most frequent therefore it is not included) is as follows: (the first digit shows the relative frequency in the group of ISN, the second in the group of CML, the third in the group PPM, the fourth in the group of the texts by R. Ivanychuk):

О — 8,15%, 8,32%, 8,20%, 8,15%; Н — 7,51%, 7,34%, 7,57%, 5,12%; А — 7,00%, 6,75%, 6,77%, 7,04%; І — 5,40%, 5,28%, 5,62%, 4,53%; Т — 5,22%, 5,24%, 5,24%, 3,94%; И — 5,11%, 5,24%, 5,04%, 5,66%; Е — 4,45%, 4,32%, 4,45%, 3,77%; Р — 4,40%, 4,10%, 4,48%, 3,65%; В — 4,22%, 4,39%, 4,22%, 5,16%; С — 3,56%, 3,75%, 3,48%, 3,64%.

The symbols with the medium frequency is presented below:

К — 2,95%, 3,22%, 3,03%, 3,03%; М — 2,90%, 2,98%, 2,84%, 2,65%; Д — 2,88%, 2,78%, 2,80%, 2,72%; У — 2,75%, 2,80%, 2,81%, 2,91%; Л — 2,61%, 2,88%, 2,63%, 3,36%; П — 2,46%, 2,37%, 2,45%, 2,48%; Я — 2,14%, 2,05%, 2,14%, 2,03%; З — 2,14%, 1,97%, 2,09%, 2,01%; Ь — 1,37%, 1,38%, 1,41%, 1,42%; Г — 1,28%, 1,23%, 1,27%, 1,51%; Б — 1,18%, 1,13%, 1,14%, 1,51%; Х —

1,17%, 1,02%, 1,18%, 0,95%; Ч — 1,01%, 1,13%, 0,94%, 1,12%; Ц — 1,01%, 0,90%, 1,13%, 0,61%.

The symbols with the lowest frequency:

Й — 0,79%, 0,79%, 0,84%, 1,42%; Ю — 0,70%, 0,76%, 0,72%, 0,76%; Ж — 0,66%, 0,68%, 0,66%, 0,75%; Ї — 0,65%, 0,66%, 0,66%, 0,49%; Є — 0,64%, 0,57%, 0,60%, 0,36%; Ф — 0,51%, 0,58%, 0,53%, 0,11%; Ш — 0,41%, 0,39%, 0,39%, 0,75%; Щ — 0,27%, 0,26%, 0,27%, 0,45%; hyphen — 0,24%, 0,26%, 0,23%, 0,07%; apostrophe — 0,15%, 0,14%, 0,15%, 0,10%; Ґ — 0,01%, 0,02%, 0,02%, 0,02%.

5 The investigation of the selection size sufficient for calculating the frequency of characters in the scientific texts

In order to determine the sample size sufficient to calculate the frequency of symbols in the scientific texts in each of the four arrays, the frequency of the characters of the extended Ukrainian alphabet was calculated. After that from the each array of data a set of selections was randomly created. Their size was from about 49% to about 69% of the size of the initial text array with a successive increase in size of the next selection to about 1.1% of the size of the text array. The upper and lower limits are chosen empirically based on the analysis of a similar study in other texts. For each selection, the frequency of the characters was counted. Based on the results, the frequencies of the symbols in the whole text and the chosen selections were analyzed as well. The theoretical background was the agreement criterion of Pearson [16] (see Fig. 1).

$$(\chi^2) \tag{1}$$

The frequency division of the symbols of the researched array was the basis for the function of the hypothetical theoretical distribution. For each selected segment, the statistics of the exponent criteria was calculated (see Fig. 2).

$$\chi^2_{\exp} \tag{2}$$

$$H_0 \tag{3}$$

For the null hypothesis (see Fig. 3) we adopted the following statement: in the selection the distribution of the frequencies of the characters of the extended Ukrainian alphabet does not differ from the corresponding distribution in the text array.

$$t_{cr} = \chi^2_{1-\alpha,k-1} \tag{4}$$

The formula (see Fig. 4) was determined according to the corresponding table [16] at the level of significance α = 0.05 and the corresponding degree of freedom k-1. The

number k (the number of characters in the extended Ukrainian alphabet is equal to 36) depended on the fact that the resulting sequences are matching the equality of minimum value in a sequence is not less than 5. If we received the result equal to the equation (see Fig. 5), then the hypothesis was rejected, otherwise it was taken.

$$\chi^2_{\exp} \geq t_{cr} \tag{5}$$

The results are presented below:

Table 3. The selection size breakdowns sufficient for calculating the frequency of characters in the scientific texts.

	ISM	CML	PPM	the texts by R. Ivanychuk
49,10%	10	9	10	10
50,20%	10	9	10	9
51,30%	9	9	10	9
52,40%	10	9	10	10
53,50%	10	9	10	10
54,60%	10	10	10	10
55,70%	10	9	10	8
56,80%	10	9	10	10
57,90%	10	10	10	10
59,00%	10	10	10	9
60,10%	10	10	10	10
61,20%	10	10	10	10
62,30%	10	10	10	10
63,40%	10	10	10	10
64,50%	10	9	10	10
65,60%	10	10	10	9
66,70%	10	10	10	10
67,80%	10	10	10	10
68,90%	10	10	10	10

As it can be seen from above, the biggest coincidence was in the PPM group that is equal to 49.1%. In the ISN group, the complete coincidence was noticed at 52.4%, in the CML group it is 64.6%, in the texts by Roman Ivanychuk it is 66.7%. Thus, in order to determine the frequency of symbols in scientific texts, the selection size of such texts in the proportion to the total amount of the texts researched, is required to be slightly smaller than the selection size of literary texts in the proportion to the total amount of the texts researched.

6 The investigation of the entropy of symbols and bigrams

The study of the entropy of symbols and bigrams was carried out as follows:

Step 1. For each of the four arrays of texts, the frequency of the characters and the numbers (from 1296 possible) and the frequency of the bigrams were calculated:

Table 4. The number of bigrams in the different text arrays.

Bigrams	ISN	CML	PPM	the texts by R. Ivanychuk
Present	1122	1141	1100	1104
Absent	174	155	196	192
The % of present bigrams	86,6%	88,0%	84,9%	85,2%

Step 2. For the received frequencies for symbols and bigrams in all four arrays the entropy was calculated according to the formula [15]:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where - p_i is the relative frequency of the i-th symbol or the bigram, N – he number of characters or bigrams.

Step 3. In each of the four arrays, we sequentially selected ten segments. The interval between them was 0.02 of the text length. We repeated steps 1 and 2 for each of them.

Step 4. In each of the four arrays we randomly selected ten segments, of the same size. We repeated steps 1 and 2 for each segment. For each segment steps 1 and 2 were repeated.

The results are presented in Tables 5 and 6.

Table 5. The entropy of the symbols.

The object of the analysis	ISN	CML	PPI	the texts by R. Ivanychuk
The whole text	4,5621	4,5577	4,5645	4,5017
The consecutive segment (the average)	4,5634	4,5582	4,5642	4,5021
The random segment (the average)	4,5636	4,5580	4,5637	4,5005

Table 6. The entropy of the bigrams.

The object of the analysis	ISN	CML	PPI	the texts by R. Ivanychuk
The whole text	8,1528	8,1619	8,1449	8,1292
The consecutive segment (the average)	8,1459	8,1537	8,1234	8,1198
The random segment (the average)	8,1440	8,1542	8,1288	8,1194

The analysis of the received symbols entropy of the segments that were selected randomly allows assuming that the entropy of symbols makes it possible to differentiate between scientific and fiction styles (see Figure 1, the first indicator on all lines of the chart is the indicator of the entropy in the entire text array). However, it is impossible to differentiate between texts themes within one stylistic group.

At the same time, the significance of the entropy of the bigrams does not allow distinguishing text by style in general (see Fig. 2, the first indicator on all lines of the chart is the indicator of the entropy in the entire text array).

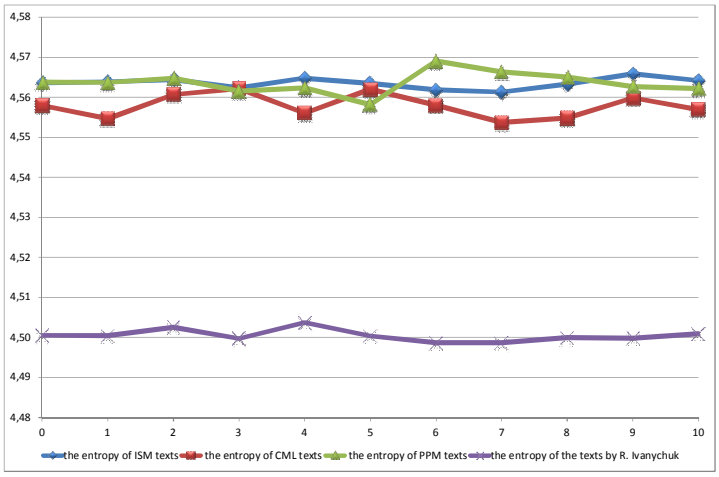


Fig. 1. The entropy of the symbols in the random segments.

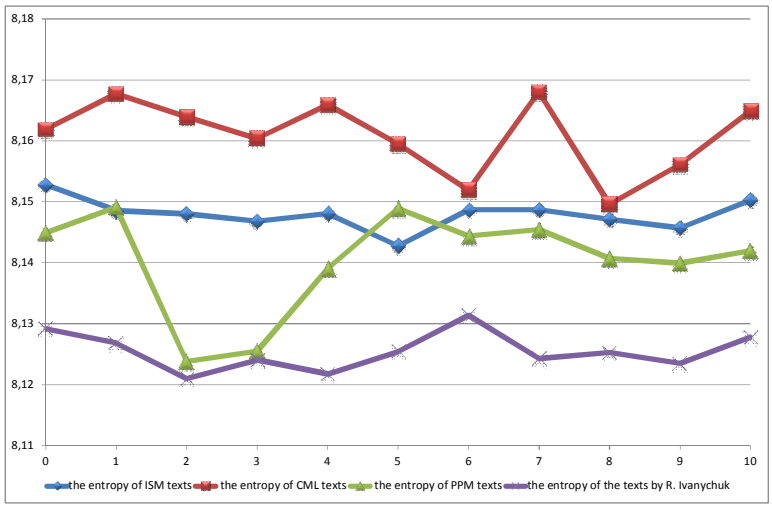


Fig. 2. The entropy of bigrams in the texts arrays.

7 Conclusions

To sum up, the statistical research of the scientific and technical texts presented in this paper is just the beginning. Due to historical circumstances, quantitative studies

of Ukrainian in general and scientific and technical in particular texts have been conducted at an inappropriate level so far. The results obtained in this research still require some clarifications. To do this it is needed to expand both the amount of research materials and the number of methods and statistical indicators. Future work concerning would include increasing the amount of research material, comparing scientific and technical texts with literary works by other authors, carrying out similar studies for three-and fourgrams of characters and adding other methods of statistical research of the language. Work must go on.

References

1. Hyland K.: The Author in the Text: Hedging Scientific Writing. Hong Kong papers in linguistics and language teaching 18, 1995, pp. 33-43.
2. Lakoff G.: Hedges: a study in meaning criteria and the logic of fuzzy concepts. Journal of Philosophical Logic, 2, 1973, pp. 458-508.
3. Light M., Xin Ying Qiu Srinivasan P.: The language of bioscience: facts, speculations and statements in between. ACL WS on Linking biological literature, ontologies and databases, 2004.
4. Vysotska, Z.: Aktualni paradyhmy doslidzhennia naukovoho. Linhvostylistychni studii 3, 43-50 (2015).
5. Visnyk natsionalnoho universytetu "Lvivska politekhnika" "Informaciini systemy ta merezhi". – Lviv (№ 673, № 689, № 699, № 715, № 743, № 770, № 783, № 805, № 814, № 829, № 832, № 854), Lviv, 2010.
6. Ivanychuk, R.: Vohnenni stovpy: tetralohiia. Folio, Kharkiv (2011).
7. Ivanychuk, R.: Voda z kameniu, Saksaul v piskakh, Manuskrypt z vulytsi Ruskoii. Dnipro, Kyiv (1981).
8. Ivanychuk, R.: Khresna proshcha : romannyi tryptykh. Piramida, Lviv (2011).
9. Ivanychuk, R.: Cherlene vyno. Kameniar, Lviv (1979).
10. Ivasenko, L.: Naukovyi styl yak obiekt doslidzhennia vitchyznianykh uchenykh. Polihraphiia i vydavnycha sprava 1(69), 111-120 (2015).
11. Kytsak, H.: Naukovyi tekst: movni ta pozamovni phaktory Mova i svit: doslidzhennia i vykladannia, eprints. zu. edu. ua/7165/1/11 ugvttm. pdf., last accessed 2018/05/15.
12. Kochan, I.: Idiostyl naukovtsia – identyphikaciina oznaka vchenoho. Rytmy suchasnoyi philolohii: zb. nauk. statei do 50-richchia prof. T.A. Kosmedy, 7–15. (2007).
13. Kulchyckyi, I.: Tekhnolohichni aspekty ukladannia korpusiv tekstiv. In: O. P. Levchenko (eds). Vydavnystvo Lvivskoii politekhniky, 2015, pp. 29-45, Lviv.
14. Kulchyckyi, I:. Korpusy tekstiv iak linhvotekhnolohichne pidgruntia vyiavlennia zmin v ukraiinskii movi. XX–XXI stolittia: zhanrovo-stylovi i linhvistychni metamorphozy v ukraiinskii movi ta literaturi. In: Arkhanhelska A. Levchenko, O. Tyshchenko, ta in (eds). Universytet im. F. Palatskoho, 2016, pp. 269-298, Olomouts.
15. Piotrovskyi, R.: Informatsyonnye izmerenyia iazyka. Nauka, Lenynhrad (1968).
16. Ruzhevych, N.: Matematychna statystyka: Navch. posibnyk dlia stud. bazovoho napriamu Prykladna matematyka. Vydavnytstvo Nacionalnoho universytety Lvivska politekhnika, Lviv (2001).

Content Analysis of Text-based Information in E-commerce Systems

Vasyl Lytvyn[1], Victoria Vysotska[2], Lyubomyr Chyrun[3], Mariya Hrendus[4], Oleh Naum[5]

[1-4]Lviv Polytechnic National University, Lviv, Ukraine
[1]Silesian University of Technologyty, Gliwice, Poland
[5]Drohobych Ivan Franko State Pedagogical University, Drohobych, Ukraine

[1]Vasyl.V.Lytvyn@lpnu.ua,
[2]Victoria.A.Vysotska@lpnu.ua,
[3]Lyubomyr.V.Chyrun@lpnu.ua,
[4]mariya.h.hirnyak@lpnu.ua,
[5]oleh.naum@gmail.com

Abstract. Instrumental means, information technologies and software for constructing such systems have been analyzed and summarized in the article based on the analysis of the basic tasks of electronic content commerce system (ECCS). ECCS functional diagram with information resources processing subsystems has been developed. The overall architecture, objectives and principles of ECCS realization were described in details. The functional elements of the system were described according to GOST 24.204.80, GOST 24.201-79, GOST 19.201-78, GOST 34.602-89, IEEE Std 1233, 1998 Edition, IEEE Std 830-1998. Software creation tools as well as management and maintenance of content were developed. The software realizations of developed ECCSs with resources processing subsystems to set up e-commerce in online newspapers and online journals were presented.

Keywords: content, commercial content, information resource, business-process, content management system, content lifecycle, Internetnewspaper, electronic content commerce system.

1 Introduction

The lack of common standardized approach to the overall ECCS design as well as process of information resources elaboration causes a number of issues while developing appropriate systems typical architecture. Due to the lack of common and detailed classification of ECCS, it becomes problematic to define and form the unified methods of information resources processing in these systems. This creates problems for the implementation of the appropriate information resources processing subsystems in ECCS such as the formation, management and maintenance of content.

2 ECCS development project

The existing ECCSs work by unknown algorithms for a wide range of programmers/specialists in the field e-business [1-4]. While creating of new ECCS the teams of specialists have to re-develop methods/information resource processing tools and content life cycle support [5-9]. Teaching and learning materials for specialists in the field are missing [10-19]. The studies concerning patterns and level of impact on the ECCS functioning relative to implement of all or some stages of commercial content life cycle for information resources processing are missing [20-28]. The analysis of ECCS functioning results aren't available because of inability to access administrative units of existing ECCS which are already known, as they are commercial projects [29-37]. The novelty of project development lies in generalized typical architecture designing as well as methods, tools and technologies for ECCS creating, and implementation of commercial content life cycle stages. Implementation of formation subsystems, management and maintenance of commercial content in ECCS leads to a reduction of production cycle and time saving while distributing commercial content, increasing of potential/constant audience and number of participants in e-business, which promotes its active development and ECCS functionality extension. The developed recommendations concerning ECCS overall typical architecture designing which differ from existing by detailed elaboration of steps and presence of sub-processing information resources that make it possible to effectively maintain content life cycle at the level of systems developer (reducing the time and resources on developing, improving the quality of system operations) [38-45]. There were developed and implemented software tools for creation, management and maintenance of content in order to reach a greater effect of operation at the level of owner (increasing profitability, growth of users interest) and user (comprehensibility, interface simplification, unification of information resources elaboration process and wider choice functional capabilities) of ECCS.

3 Expected effects of ECCS introduction

Expected economic effect from ECCS application is predicted by reducing expenses for project development and architecture of the system, for additional resources and personnel usage in the absence of clearly regulated plan. The factors of economic impact are following.

1. The presence of commercial content forming subsystem reduces expenses for informational resource moderators extra staff, as it carries out the authors' and moderators' work, namely: collecting data from various sources, formatting and categorization of content, keywords and duplication identifying, digest formation as well as content distribution.

2. Content management subsystem reduces expenses for personnel responsible for information resource updating.

3. Content support subsystem reduces expenses for personnel responsible for collecting and system functioning analyzing.

4. The presence of information resources processing subsystems allows reducing time on actual unique commercial content prompt getting, which leads to ECCS target audience increase, and, respectively, increases ECCS implementation economic effect on several positions.

ECCS application leads to the growth of productivity of the project participants, improving the quality of information resources processing, reducing time spent on ECCS implementation and actual unique commercial content prompt receipt. The reasons for effect growth are following.

1. Increase of labor efficiency occurs through the usage of the results of additional professional resources work, such as Google Analytics, moderators, administrators, programmers and analysts.

2. Analysis and increase of labor efficiency leads to improving the quality of formatting, management and maintenance of the content as well as reducing time spent on system implementation and prompt content receipt.

3. Improving the quality of information resources processing is based on the analysis of statistics and main characteristics of ECCS functioning, such as number of visits, average time spent on visit, failures indicator, achievement of search goal, content dynamics, number of page views, number of page views per visit, new visits, absolutely new visitors, traffic source etc.

4. Reducing time spent on ECCS implementation and actual unique commercial content prompt receipt improves the decision-making quality for e-business participants:

a) The authors in order to create actual content according to the established, selected and distributed the list of digests related to the actual subject;

b) The moderators in order to form rules as well as operational data collecting source lists;

c) The moderators in order to form rules for categorization, duplication, formatting and content management, resource formation;

d) The administrators in order to perform resource and system management;

e) The analysts in order to perform the research system functioning statistics, formation of rules for identifying new stories and personalization of cooperation with user, content ranking.

Organizing effect lies in reducing the number of staff (moderator 1-3, administrators 1-2, authors 1-10, analysts 1-3, programmers 1-2), which is launched to preparation, formation and decision-making and changing its functions, which are performed by formation, management and content support subsystems (data preparation data for authors, tracking the results of staff performance, data collecting for analysts and moderators); changing the organizational structure (a clear division of functions between project participants, i. e. moderator won't be analyst, and vice versa); reducing the number of operations performed by personnel (a part of operations should be performed by ECCS via information resources processing).

Technological effect through the release or reducing such resources as staff, and more efficient usage of information resources processing subprograms in ECCS as well as clear allocation of responsibilities among participants, launched to certain project. The development of the new technologies such as formation, management and maintenance process of the content, and search engines peculiarities.

Ergonomic effect lies into influence of the results of ECCS operation and information resources processing through the formation, management and maintenance of content on number of visits, average time spent on information resources (min.: s), failures indicator (%,) achievement of searching purpose, content dynamics (%), total number of page views, number of page views per one visit, new visits (%), absolutely new visitors, traffic source in %, etc.

Psychological effect lies in friendly interactive interface for each participant of the project, which simplifies work for authors, moderators, administrators, analysts, and improves psychological indicators for visitors and regular ECCS users through individualization of work.

*Advertising effec*t lies in unique content, content template and information resource usage, which improve searching results from search engines, and serve as self-promotion of ECCS, information resource and content. Cooperation with Google advertisement improves ECCS information resource advertisement rates as well as e-business.

Social effect lies in increasing the number of information resource visitors (permanent and unique), increasing the limits of target audience, resource availability and commercial content support, coverage of a wider social audience, ability to change the boundaries of the target audience through social regulation topics and information resource filling. The support of topically similar and relevant content, its uniqueness and efficiency of creation, formation, representation through information resource and support allows adjusting the limits of constant target social audience for ECCS and forecasting/adjusting these limits changes.

4 ECCS input data

List of input data types, their characteristics, description, classification allow them be conditionally divided into seven groups of incoming content, depending on the project participants' class: visitors, users, authors, administrators, moderators, analysts and other resources. The input data elements structure allows you to create the requirements to ECCS and its components as well as clearly limit their functionalities, describe the source, their incoming frequency and additional conditions/limitations which are imposed by input data source. ECCS input data are following:

1) a content from different sources specified by moderator predefined list (subscription, free and open content from information sources predefined list, author content, content as the result of the search engines operations etc.) in the form of data without predefined structure in HTML/XML-format in order to create commercial content [1-4];

2) information requests from users/visitors ECCS information resource as a text message in the appropriate fields (keywords/phrases in order to find commercial content, messages and/or requests to participants of the project) [1];

3) actual data (set of source address, words/phrases, and/or user logins with disabling printing/access option) and/or rules (set of operations like IF (fact i) THEN ban ELSE check OR permit/print) from ECCS information resource moderators;

4) actual data (ECCS and Google Analytics parallel cooperation statistics as XML-tables and/or fixed customized/personalized user actions such as viewing, downloading, storing and/or content searching) and/or statistical analysis rules of user action (formation of associative lists of popular, topical, outdated, recently revised and/or similar content/subjects or author works, and forecasting of thematic content demand) from analysts of ECCS operation;

5) information resources URL-address from moderators for information filters databases such as ECCS data source [1];

6) language dictionaries replenishment by moderators as a list of words, phrases with the defined characteristics (part of language and gender, number, cases etc.) as well as additional set of morphological features to each of them [3-4];

7) ECCS information resource operation statistical data, collected at specified intervals from Google Analytics as XML-tables [3-4];

8) comments and user feedbacks as a text data array completed in specially designated places of information resources;

9) members voting results regarding content and quality of the commercial content according to the defined numerical evaluation scale with the possibility to support linguistically inaccurate voting, e. g. “good” etc.;

10) individual statistics/personalized user actions (viewing of content, viewing time, content downloading/uploading, searching);

11) database content components, content collected from various sources, duplicate content, registered users, project participants, linguistic dictionaries in order to determine keywords and headings, keywords and subjects, keywords in order to determine subjects and their replenishment;

12) external advertisement in the form of banners from Google and partner sites;

13) thematic stickers of information and/or entertainment content (weather, exchange rates, anecdotes, announcements, etc.);

14) ECCS settings from administrators in the form of change/creation and liquidations of additional options and system configurations and/or information resource through the administrative part of the system.

5 ECCS output data description

ECCS output data are following [1-4]:

1) the final information product of ECCS operation in the form of content (article, announcement, digest, e-book, audio, video etc.);

2) answers on user's information requests (list and set of a similar subject content according to the content searching results using keywords and concepts which contained in the information request);

3) commercial content digests (short thematic announcements of content according to their set of keywords and concepts);

4) information resource visits' statistics according to the number of clicks on the link, time spent on visit, number of page downloads and additional investments;

5) user actions' statistics and/or resource visitors in order to form individual portrait of user/audience according to the number of clicks on the link, time spent on visit, number of page downloads and additional investments, switching between pages etc.;

6) forming/filling in of information resource page individually for user according to its statistical action history (selection of related content, current recommendations for thematic content etc.);

7) new rubrics/topics of commercial content (creation and formation of a new category according to the input data sources analysis, authors operations, comments and requests from users);

8) ranking the results of commercial content in the form of evaluation within a scale [0;5], [0;10], [0;12] or [0;100];

9) relationship tables of similar, recently revised, popular, outdated, authorial and/or topical content according to the keyword list of this content, user pages viewing analysis, sequence and time spent on thematic content viewing;

10) comments evaluation as the result of the content user's comments in the form of permission/ prohibition of printing on resources, and, if needed, with the prohibitive recommendations for certain user to post the following comments.

The list of output data, messages, their description, characterization, classification, method of formation and transfer allows to create precise functional requirements for the development and implementation of ECCS.

6 ECCS functional requirements

ECCS interface basic requirements: scalability/ performance while working with a large number of users, sessions, transactions and database connections; productive browser connection and back-end data storage; rapid development and deployment of Web OLTP-applications support; support of synchronous/asynchronous transaction management via servers. ECCS transactions servers characterize the following features: built-in transaction management services; the mechanism of starting and management of servlets; calls of distributed objects in order to ensure communication in multilevel applications; facilities for rapid software development for intermediate level, including component development. ECCS provides support to six interfaces: with limited access for visitors and users; without restrictions for the administrator and moderator; with free access to author and analyst. The access for users is implemented using login and password. Such services as a choice of content for a

certain period of time starting from the beginning of content filling through a calendar are additionally implemented there. The convenient rubricator allows you to select content by certain category. Searching is carried out by using keywords in the database.

System administration is carried out through administrator interface access to which is limited and implemented using login and password. The adjustments to the structure of the system/resources are made here as well as user access rights are added, edited or deleted, and content distribution rule change. Creating/editing content is carried out through author interface access to which is limited and implemented using login and password. The development of functional requirements for formation subsystems, management and maintenance of content in ECCS promotes the development of such systems typical architecture. On the one hand, ECCSs facilitate work of moderators, authors, analysts and administrators of these systems, and increase system functionality to their users on the other.

ECCSs select a range of topical issues in the form of content plurality from other sources for moderators and authors according to their rating through commercial content formation subsystem. The author creates his own commercial content according to ECCS analysis chosen from various sources of actual content. If needed, the moderator creates new rules for filtering content from various sources and updates the addresses of other sources in commercial content formation subsystem. Analyst analyzes the activity of the target audience and ECCS operation for the development of new rules and statistical analysis of the dynamics of commercial content lifecycle stages through commercial content support subsystem. These rules should increase the range of the target audience, number of visits, number of unique visitors, number of repeated visits, number of visits from search engines, number of direct visits, number of regional visits, number of visits etc. for thematic information resources in ECCS. The purpose of work is also to determine the functional requirements for information resources processing subsystems in ECCS as formation, management and maintenance of commercial content. Commercial content formation subsystem facilitates the work of the authors and moderators of ECCS. Content management subsystem facilitates the work of the authors and moderators of ECCS as well as supports different functionalities for users of these systems. Commercial content support subsystem facilitates the work of the analysts of ECCS. Information resource moderation is carried out through moderator's interface access to which is limited and implemented using login and password. Here are introduced the rules and parameters for monitoring content from different sources; commercial content is being added, edited or deleted; the content of the day is set (for publishing ECCSs); the content of static pages as well as rules of sending letters with the content is changed. There was implemented a convenient service for adding new clients with a possibility to group them into certain categories and set the time limits of access that is blocked automatically after a certain period. Statistic formation and its analysis is carried out through author interface access to which is limited and implemented using login and password. Here are made the amendments in the rules of information resource operating statistics; this statistics analysis rules are being added, edited or deleted; the

rules of commercial content ranking as well as commercial content rubrics and commercial content authors are established. There was implemented a convenient service for automated adding of information resource operating statistics and this resource users activity as well as the rules of this statistics analysis.

Software tools created for content version control ensure that the online portals content will not be lost or accidentally overwritten. Moderators and administrators have the opportunity to easily find the required version of content and information resource. Building of business processes based on roles and user groups means the independence from delays while execution by individual persons.

Structural elements designed for constructing interfaces and information resources processing software tools [1-4]:

1) Communication protocols (HTTP, FTP, IIOP).

2) Integration of HTTP-servers with information sources (CGI, Perl, PHP and specialized API).

3) Implementation of hypertext capabilities (HTML, WML, XML, XHTML, JavaScript).

4) Implementation of multimedia capabilities (Flash, formats for audio/video, VRML).

5) Communication implementation (POP, UDP, SMTP).

6) Calculation support (PHP, Java).

7) Systems and service of content management (CMS, CMIS).

8) Network management protocol (CMIP).

9) Organization of mobile access and calculations (GPRS, EDGE,UMTS, WAP).

10) Implementation and development of distributed objects (CORBA, COM,DCOM, EML, ORB).

11) Saving and processing of data (File systems, OS, database management systems).

ECCS technology is automation (full or partial) of the business process by which content, documents, information or tasks are transferred for the appropriate action from one participant to another according to the set of procedural rules. ECCS describes, creates and manages workflow (business process) using the software that interprets the process description, cooperates with the participants of the workflow and, if needed, elicits corresponding software applications and instrumental tools. ECCS automates a business process (not a function), and implements the rules of interaction between participants, as these aspects are the main centres of losses because of their ambiguity. The result of ECCS formation is such a system like online newspaper, online magazine, online editions, online publishing, distance learning, online store created for selling content in the form of e-books, photo, video, audio etc.

Standardization and implementation of functional requirements for ECCS formation provides the creation of a generalized approach for such systems developers in order to reduce the time for formation and implementation of such systems with avoiding of appropriate project development phase. Requirements for ECCS operation results, operation regulation, ways of displaying, transfer and store depends on the implementation of major information resources processing subsystems

such as formation, management and maintenance of commercial content. Requirements for compatibility and ways of interacting and communication with other systems lie in support of text arrays processing in HTML and/or XML-format. ECCS ergonomic requirements lie in convenience of maintenance and system maintenance, rational configuration of program and interface elements, convenience of system management tools, aesthetic design. ECCSs support security/protection of data and other system components from unauthorized access, loss, destruction, damage. Support of organizational and procedural requirements for personnel, its composition and qualifications, system operation charts, rights and powers to operate the system etc. allows to implement/introduce ECCS, maintain the system functioning process on high level and analyze the results of its approbation and its major subsystems of information resources processing.

The commercial content formation subsystem is based on the multilevel model of processes organization. Such an organization provides the separation of subsystem structure on such individual modules: gathering/creation of content from various sources, formatting, identifying keywords and concepts, headings, identifying duplication, formation of digests and selective dissemination of content between ECCS users. Content formation subsystem is implemented in accordance with the algorithm 1-2.

Algorithm 1. Formation of content analysis set

1st. Gathering content from various data sources and saving it in DB.

2st. Filtering content by a set of rules defined by system moderator.

3st. Formatting content in XML-format and saving it in the database.

4st. Determination of content duplication and duplicates filtering.

5st. Determination of keywords and concepts of the content and saving them in the database.

6st. Formation of content digest and saving it in the database.

7st. Content formation and making an entry in the annotated DB.

8st. Commercial content categorization.

Step 1. Identification of concepts from content using data from definitions database.

2sp. In case experts weights are presented in the definition of the concept, then it should be calculated the weight of concept emergence in the text, taking into account the frequency of the phrase appearing in content.

3sp. Adoption of decision on content belonging to a particular category based on the rules of categorization and set of concepts found in the text of the calculated weights.

9st. Selective content distribution among moderators and authors of commercial content according to the direction of their work and their estimated rating of popularity and quality of work.

The principle of identifying keywords within the meaning (terms) based on Zipf's law and it comes down to choice of words with average occurrence frequency (the most used words are ignored due to “stop dictionaries”, and the text rare words aren’t taken into account). Then the terms for new keyword formation are synthesized, using

structural parts of speech base. The process of categorization using automatic indexing of commercial content components is divided into consecutive blocks: morphological analysis, syntactic analysis, semantic and syntactic analysis of linguistic structures and text content substantial writing variation. Based on analysis of the given set of content, authors create commercial content, which then passes the following stages of processing (alg. 2.)

Algorithm 2. Content formation for information resource in ECCS

1st. Author analysis of content plurality obtained from various sources and filtered.

2st. Author content formation as a result of the analysis of content plurality obtained from various sources and filtered.

3st. Author content formatting in XML-format and saving it in DB.

4st. Checking for commercial content duplication and doubles prohibition. In case there are duplicates, the author's bugs should be fixed, and move to the 10st, otherwise - move to the 5st.

5st. Verification and validation of commercial content. In case there is unique content > *w*, move to the 6st, otherwise - the author's bugs should be fixed, and move to the 10st.

6st. Determination of keywords and concepts of the content and saving them in the database.

7st. Author commercial content categorization.

8st. Formation of content digest and saving it in the database.

9st. Content formation and making an entry in annotated database.

10st. Recalculation of commercial content author rating.

Content distribution subsystem selectively sends digests among the authors according to their work quality rating. From the very beginning the subsystem receives ready digests from sources via RSS. Then digests are distributed among authors according to their rating: The author with the highest rating is the first person who receives digests for revision. In case all operations are ready, the subsystem goes into standby mode before the appearance of new content. Authors' rating indicates the performance/effectiveness of work of each of them individually. It is affected by the next criteria: the percentage of unique content (quality of the author work), the number of content views (weight of search and direct conversions), user evaluation (users' activity) and time spent on visit (a measure of users' interest in the content). Static representation of the content support model that describes the attributes and behavior of subsystems is presented on a class diagram. Principles of content analysis are the basis for the implementation of the subsystem. This allows automating of the various stages of formation of information product of such a type without loss of content and reducing quality. The results of its application while developing a number of commercial content projects confirm the effectiveness of subsystem functioning. The developed automation means for content formation allow accelerating of content formation process and increasing the usage rate of information resources which were created by their help. Commercial content management subsystem supports interactive communication between user and electronic content commerce system through information resource. The subsystem must form information resource to the

needs user and respond to his/her requests. Information resource consists of a clearly defined set of components (nodes) that are semi ordered (some of the components are clearly defined order, and their location in the graph is not necessarily defined). The number of information resource components and the corresponding volume of components set is precisely defined/evaluated. Construction of graph of navigation (navigation graph) on information resource is going on the basis of given ratio of adherence to the set of relations of resource components. Navigation is carried out without interruption and transition to a new node of the navigation graph is logical. The units are submitted by the random types of relationships that do not have significant limitations. The minimum number of project nodes determines the graph that provides the most meaningful information about information resource. Each arc in the navigation graph displays the elementary relationship between components of resource and must have an orientation. In order to receive content with certain parameters the user must pass a number of steps. Content management process by user or moderator is implemented through alg. 3.

Algorithm 3. Content management.

1st. Authorization in electronic content commercial system.

2st. Selecting a template content from the dialog box.

3st. Selecting and setting various parameters.

4st. Selecting the required functional under the protocol template.

5st. Content generation out of template by the 3st defined criteria.

The basis of content management subsystem is a core - a subsystem that connects all parts into a single application and is responsible for loading and configuring modules, connecting general dependencies and providing information resources integration points. The most important task is to ensure that the information needs of system problem-oriented elements, maintaining access to data of different categories of users, compliance integrity and consistency of data, minimization and control under data excess, ability to develop and change the internal organization of information resource, compliance with the requirements of quality and data efficiency. ECCS provides modification of information resources through submission methods, formats, and content internal organization; content storing environment, physical storage units, technical means; user requirements, the emergence of new requirements and types of users; procedure for distribution of content and methods of user access.

7 Conclusion

Content analysis enables formation of conclusions regarding the possibility of influence on target audience growth, tendencies of information resource annual volume of visits increase over time. It is possible to investigate the dynamics of changes in global, national, regional, industrial, thematic content flow or dynamics of growth of homogeneous flow part on any basis. The basis for the obtaining data on dynamics of change and target audience growth is secondary sources, such as, Google

Analytics. In order to improve the accuracy of content analysis dynamic results there should be implemented the following steps: users' information portraits and content analysis, analysis of content thematic subjects, formation of content relationship tables, comments tonality, rating/content history and information resource statistics analysis.

References

1. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
2. Content Management Interoperability Services, Version V. 0.5, Hopkinton: EMC (2008)
3. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 47–55 (2017)
4. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)
5. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Germany: LAP LAMBERT Academic Publishing (2016).
6. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
7. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
8. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
9. CM Lifecycle Poster, Content Management Professionals, Access:http://www.cmprosold.org/resources/poster/ (2010)
10. Hackos, J.: Content Management for Dynamic Web Delivery. Hoboken: Wiley (2002)
11. Halvorson, K.: Content Strategy for the Web. Reading: New Riders Press (2009)
12. McGovern, G., Norton, R.: Content Critical. Upper Saddle River: FT Press (2001)
13. McKeever, S.: Understanding Web content management systems: evolution, lifecycle and market. In: Industrial Management & Data Systems (MCB UP), 103(9), 686–692 (2003)
14. Nakano, R.: Web content management: a collaborative approach. Boston: Addison Wesley Prof. (2002)
15. Papka, R.: On-line News Event Detection, Clustering, and Tracking: thesis for the degree doctor of philosophy. Amherst: Massachusetts University (1999)
16. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
17. Rockley, A.: Managing Enterprise Content: A Unified Content Strategy. New Riders Press (2002)
18. Stone, W. Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! In: IEEE Antennas and Propagation, 45(4) 47-49 (2003)
19. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)

20. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
21. The Content Management Possibilities Poster. http://metatorial.com/pagea.asp?id=poster.
22. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
23. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. . In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 100–109 (2017)
24. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)
25. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
26. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
27. Hamon, T.: Biomedical text mining. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf (2017)
28. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)
29. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
30. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
31. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)
32. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
33. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf. (2017)
34. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
35. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)

36. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
37. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 133–134 (2017)
38. Kirkin, S., Melnyk, K.: Intelligent data processing in creating targeted advertising. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 131–132 (2017)
39. Hordienko, H., Ilchenko, M.: Development and computerization of an English term system in the fields of drilling and drilling rigs. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 129–130 (2017)
40. Gorbachov, V., Cherednichenko, O.: Improving communication in enterprise solutions: challenges and opportunities. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 127–128 (2017)
41. Didusov, V., Kochueva, Z.: Statistical methods usage of descriptive statistics in corpus linguistic. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 125–126 (2017)
42. Verbinenko, Yu.: Discursive units in scientific texts. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 120–123 (2017)
43. Titova, V., Gnatchuk, I.: Evaluation ofa formalized model for classification of emergency situations. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 110–119 (2017)
44. Kotov, M.: NLP resources for a rare language morphological analyzer: danish case. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 31–36 (2017)
45. Kuprianov, Ye.: Semantic state superpositions and their treatment in virtual lexicographic laboratory for spanish language dictionary. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 37–46 (2017)

PART II.
WORKSHOP CONFERENCE TRACKS

Section I.
Computational Linguistics

COLINS'2018, Volume II: Workshop. Lviv, Ukraine, June 25-27, 2018, ISSN 2523-4013
http://colins.in.ua, online

Detection of Gaps in Documentation Concerning Remote-piloted Aviation based on Content Analysis

Volodymyr Markiv[1], Mykola Mykyichuk[2], Oksana Markiv[3]

Lviv Polytechnic National University, Lviv, Ukraine

[1]vovamarkiv6230@gmail.com, [2]mykolamm@ukr.net, [3]markivoksana89@gmail.com

Abstract. This article dwells upon the procedures of computer-linguistic formation of documentation concerning remote-piloted vehicles that are necessary for the implementation of software components for the collection, processing and preservation of information received from open web-resources. It is highlighted that processing of large volumes of information located on open web-resources requires development of certain approaches to its automated consolidation, classification, structuring, processing and use that is very useful for further certification of remote-piloted vehicles. It is emphasized that linguistic support as a set of linguistic means of information processing is important during development of relevant information systems for aviation support. The procedures for detecting gaps in documentation concerning remote-piloted aviation by analyzing the structure of documentation, its content, frequency analysis of markers and indicative features that allow to improve functional quality indicators are very similar to the same procedures concerning tourism documentation. It has been made based on documentation structure and content analysis and is important for automatic system development of documentation creation in the sphere of remote-piloted aviation.

Keywords: Remote-piloted vehicles, Documentation, Information gap, Indicative characteristics, Linguistic markers, Remote-piloted vehicles certification

8 Introduction

At the present stage of market economy development remote-piloted aviation is one of the promising areas both in the internal and external economic activity of the country and its regions. Intensive development of remote-piloted vehicles use in terms of present days causes growth of the knowledge base and the number of remote-piloted activities subjects. It is therefore necessary to provide the product consumer with the necessary, accurate and truthful information through the formation of a complete information resource in the form of so-called quality documentation.

The need of product consumer to obtain full information about remote-piloted aviation, as well as the increased tendency in the number of consumers that

independently search for information using the Internet by increased discussion of services on the forums cause the necessity the take into consideration needs of product consumer.

This can be achieved through continuous interaction between the producer and the consumer of the remote-piloted vehicle through the data collection about service and the formation of quality documentation. The need continuously to improve the quality of aviation service cause the necessity to analyze in details the documentation and use an appropriate method of evaluation of its quality based on the needs of the product consumer.

The formation of remote-piloted vehicle documentation (RPD) is a complex process that requires time consuming, highly qualified specialists, using of information technology for information processing, its grouping, identification of information sources, logical analysis of materials, choice of methods of data extraction, data aggregation to single accessible to the consumer product format by consolidating the processed information. Effective formation of such a documentation is practically impossible without the development of appropriate components of software and algorithmic complex of remote-piloted aviation support [3, 4].

In the conditions of intensive globalization and significant development of information society, there is a rapid increase in the number of services and the rapid development of the remote-piloted vehicle industry. At the same time, the quantity of information resources with open access, which contains information on remote-piloted vehicles, is significantly increasing.

It is necessary to create appropriate conditions for the prompt processing of all available information in order to bring it to a single standard view, in order to be able to process and use information in the future depending on the needs of services. It is needed to take into account both the processes of data collection and the processes of their verification, cleaning and transformation for further integration with the use of different methods and tools.

9 Development of procedures for detecting gaps in documentation concerning remote-piloted aviation

The legal status of remote-piloted vehicles is currently quite uncertain in Ukraine. Although the Air Code of Ukraine defines the concept of an "unmanned aerial vehicle", the basic regulatory documents regulating the use of Ukraine airspace do not mention the remote-piloted vehicles.

The use of unmanned aerial vehicles, as well as any kind of activity that may be potentially dangerous, should obviously have some legal basis. However, a change in legislation is a rather inertial, slow process that does not keep up with the rapid development of technologies. So today, not only in Ukraine, but also in other countries, the legislative regulation of activities related to the use of remote-piloted vehicles is at the stage of becoming topical. Some changes in legislation are constantly taking place, and rules and norms are being adopted and amended.

There are currently no specific regulatory documents on the performance of unmanned aerial vehicles in Ukraine. There is no complete legal basis for certifying remote-piloted vehicles, licensing crews that control them, and concerning with the safety of remote-piloted vehicles, even in temporarily reserved airspace. The national regulatory framework does not contain provisions on the integration of unmanned aerial vehicles into the air traffic management. Ukraine is ready to cooperate with other European institutions and all interested departments to develop joint efforts of the appropriate level of regulatory acts that will regulate the use of unmanned aircraft in Ukraine[7, 8].

The use of drones in Ukraine must be managed and controlled. All these requirements are not a new fiction of the State Aviation Service, it is its perception and transmitting into Ukrainian realities of the experience of the leading countries of the world, primarily the United States and Great Britain. But there are a lot of incorrect information concerning the remote-piloted vehicle aviation. It is necessary to detect and illuminate such gaps in content.

The procedures for detecting gaps in documentation concerning remote-piloted aviation by analyzing the structure of documentation, its content, frequency analysis of markers and indicative features that allow to improve functional quality indicators is very similar to the same procedures concerning tourism documentation.

Developed algorithms for comparing atomic situations of web-page posts with elements of documentation allow to fill it with consolidated information from the open web-resources. However, these algorithms do not indicate the possible inaccuracy and incoherence of both the structure of the tourism documentation, and its content in the form of facts and events. One of the methods, which allows to find and eliminate the shortcomings of documentation, is a method of detecting and filling gaps[7, 8, 10, 11].

The notion of "gap" is used in many fields, in particular, in the legislation the gap is the lack of the required norm in legislative acts. In this article, under the gaps in the tourism documentation has been understood certain part of the tourism documentation that is has not full of information about the facts or events [1,2].

The general tourism documentation model allows to distinguish two types of gaps [6, 16, 21-51]:

1. structural gaps in the tourism documentation;
2. gaps in the content of tourism documentation.

Such types of gaps also are used in the remote-piloted vehicle documentation

2.1. Peculiarities of identifying structural gaps in documentation

So, structural gaps in RPD may arise when the structure of the RPD does not correspond to the predicted requirements, which are formulated both by the supplier and the consumer of the product.

In general, the presence of structural gap can be formulated as follows: the set of the same type elements of RPD contains fewer elements than the corresponding set of these elements in the requirements of the RPD. That is, if $Element^{(RPD)}$ and $Element^{(Requirement)}$ respectively, the set of elements in the RPD and in the requirements of the RPD, then in the general case there is a structural gap, if:

$$Element^{(Gap)} = Element^{(Requirement)} \setminus Element^{(RPD)} \neq \varnothing \quad (1)$$

The elements of RPD, which can cause structural gaps, are remote-piloted aviation objects (O), remote-piloted aviation objects actions (OA), the classifiers and the terms of the classifiers. To eliminate these gaps it is necessary to supplement the corresponding sets with new elements.

The supplement of the O and OA sets of new elements takes place according to algorithm with the following steps[1, 7, 8, 13]:

1) add a new element to the *Object* set according to the requirements of the RPD, specifying the attributes of the fact of the definition of this object including classifiers;

2) determine the indicative feature of the object ;

3) include the object $Object_i^{(Gap)}$ in services and consumer reviews in accordance with the requirements of the RPD;

4) add a new element to the set of actions $Action\left(Object_i^{(Gap)}\right)$ in accordance with the requirements of the RPD, specifying the attributes of the determination of this action for , including classifiers;

5) determine the indicative trait of the OA $ActionIndicator\left(Action\left(Object_i^{(Gap)}\right)_j\right)$;

6) include OA $Action\left(Object_i^{(Gap)}\right)_j$ in services and consumer responses in accordance with the requirements of the RPD;

Note that the procedure for adding a new action to an existing object in the RPD should begin with item 4 [14, 15].

The addition of sets of classifiers and terms of classifiers of RPD with new elements takes place by the algorithm in the following way:

1) add a new element $Classifier_i^{(Gap)}$ to the *Classifier* set in accordance with the requirements of the RPD;

2) add a new element $Term_{ij}^{(Gap)}$ to the set of terms $Term_i^{(Gap)}$ of the classifier $Classifier_i^{(Gap)}$ in accordance with the requirements of the RPD ;

3) to the plurality of links between terms $TermRels_i^{(Gap)}$ add links of the new element $Term_{ij}^{(Gap)}$ with other terms of the classifier $Classifier_i^{(Gap)}$;

4) determine the indicative sign of the term $TermIndicator\left(Term_{ij}^{(Gap)}\right)$.

Note that the procedure for adding a new term for the existing RPD classifier should begin with item 2.

2.2. Peculiarities of detecting content gaps in documentation

Gaps in content in the RPD arise because of the lack of information on the facts and events associated with certain structural elements of the RPD - O, OA and the terms of the classifiers.

In general, the presence of content gap can be formulated as follows: for structure element of RPD not all necessary facts and events are associated with it. There are the following types of content gaps [5, 12]:

1. facts and facts of the definition of O and OA are not associated with the terms of the classifiers;
2. facts are not associated with O and OA;
3. events are not associated with O and OA;

If $Term_{ij}$ - the term of the classifier, $ElementDefiniton^{(TD)}(Term_{ij})$ - the set of all the facts of the definition of O and OA associated with this term in the RPD, $ElementDefiniton^{(Expected)}(Term_{ij})$- the set of all facts of the definition of O and OA, which should be related to this term, then there is content gap by classification of the facts of determination, if:

$$ElementDefiniton^{(Gap)}(Term_{ij}) = ElementDefiniton^{(Expected)}(Term_{ij}) \setminus ElementDefiniton^{(RPD)}(Term_{ij}) \neq \varnothing \quad (2)$$

If $ElementFact^{(RPD)}(Term_{ij})$ - the set of all O and OA facts related to the term in the RPD, $ElementFact^{(Expected)}(Term_{ij})$ - the set of all O and OA facts that would be expected to be related to this term, then there is a gap in the content of the classification of the facts if:

$$ElementFact^{(Gap)}(Term_{ij}) = ElementFact^{(Expected)}(Term_{ij}) \setminus ElementFact^{(RPD)}(Term_{ij}) \neq \varnothing \quad (3)$$

If $ElementEvent^{(RPD)}(Term_{ij})$ – set of all events of O and OA, related to term $Term_{ij}$ in RPD, $ElementEvent^{(Expected)}(Term_{ij})$ – set of all events of O and OA, that may be related to the term $Term_{ij}$ in RPD, then there is content gap by event classification, if:

$$ElementEvent^{(Gap)}(Term_{ij}) = ElementEvent^{(Expected)}(Term_{ij}) \setminus ElementEvent^{(RPD)}(Term_{ij}) \neq \varnothing \quad (4)$$

If $Element_i$ – O or OA, $ElementFact^{(RPD)}(Element_i)$ – set of all facts of O or OA in RPD, $ElementFact^{(Expected)}(Element_i)$ – expected set of all facts of O or OA, then there is fact content gap, if:

$$ElementFact^{(Gap)}(Element_i) = ElementFact^{(Expected)}(Element_i) \setminus ElementFact^{(RPD)}(Element_i) \neq \varnothing \quad (5)$$

If $ElementEvent^{(RPD)}(Element_i)$ – set of all O or OA $Element_i$ in RPD, $ElementEvent^{(Expected)}(Element_i)$ – espected set of all events in O or OA, then there is the event content gap, if:

$$ElementEvent^{(Gap)}(Element_i) = ElementEvent^{(Expected)}(Element_i) \setminus ElementEvent^{(RPD)}(Element_i) \neq \varnothing \quad (6)$$

Gaps in the content according to classification of the facts of definition (2) is eliminated by the expert, which should determine the plurality of terms that are associated with O or OA.

Gaps in the content according to classification of facts and events (3), (4), and the gaps in the content of facts and events (5), (6) are eliminated in the following ways:

1) use of new open web-resources that are relevant to the available indicative characteristics;

2) specification of the indicative characteristics of the O, experience, time component, OA and the classifier term by the frequency analysis method.

To identify new open web-resources that are better suited to the existing indicative traits, it is necessary to apply the algorithm for forming a set of web-resources according to keywords that are specified in markers of indicative signs, and check the usefulness of web-pages of these open web-resources by criterion [7-11, 22-23].

2.3. Frequency analysis of markers and indicative characteristics to eliminate gaps in content

Frequency analysis is widely used in cryptography and is based on the explicit of statistical distribution of individual characters and their sequences in the text of the message. That is, frequency analysis suggests that the frequency of occurrence of a given character in rather long texts is one and the same for separate texts of messages.

If *Element* is a certain element of the RPD, which reflects either a certain O, OA or the term of the classifier, or a certain sign of experience or time component, then, accordingly, on the basis of the indicative characteristics of the O, OA, the classifier term, experience, time component, generalize the concept of the indicative characteristic for *Element*:

$$ElementInd\,icator = \left\{ \left\langle \begin{matrix} Marker\,(ElementInd\,icator\,)_i, \\ \mu(ElementInd\,icator\,)_i \end{matrix} \right\rangle \right\}_{i=1}^{N^{(ElementIndicator)}} \quad (7)$$

where $Marker(ElementIndicator)_i \in Marker$ - the i-th marker to display the RPD element, $\mu(ElementIndicator)_i \in [0,1]$ - the measure of correspondence of the i-th marker to the RPD element, $N^{(ElementIndicator)}$ - the number of markers that reflect the element of the RPD.

In order to use marker frequency analysis to eliminate content gaps in the RPD, it is necessary to analyze the frequency of occurrence of the corresponding markers when comparing the atomic situation with the components of the documentation.

If in the result of the open web-resources processing, $N^{(PostPart,Processing)}$ atomic situations have been detected, and $N^{(PostPart,Marker(ElementIndicator)_i)}$ of them are in comparison with the indicative characters containing the marker $Marker\,(ElementInd\,icator\,)_i$, then the appearance frequency of marker $Marker\,(ElementInd\,icator\,)_i$ in atomic situations is the following:

$$MarkerFrequency(Marker(ElementIndicator)_i) = \frac{N^{(PostPart,Marker(ElementIndicator)_i)}}{N^{(PostPart,Processing)}} \quad (8)$$

To evaluate the frequency of the marker the following criterion is introduced:

$$MarkerFrequency(Marker(ElementIndicator)_i) \geq \alpha^{(MarkerFrequency)}, \quad (9)$$

where $\alpha^{(MarkerFrequency)} \in (0,1]$ - the minimum allowable value for the frequency of the marker.

The presence of many markers that do not meet criterion (9) may be due to the presence of gaps in content. Gaps on the basis of frequency analysis of markers are eliminated in the following ways [3, 16, 18]:

1) use of new open web-resources that better suit to markers that do not satisfy the criterion (9);

2) reassessment of the marker compliance $\mu(ElementInd\,icator\,)_i$ of marker $Marker(ElementInd\,icator\,)_i$ for each indicative characteristic (7) where this marker occurs;

3) remove of the marker $Marker(ElementIndicator)_i$ from a set of all markers *Marker*, and completing of this set with new markers.

The use of the frequency analysis of indicative features to eliminate content gaps in the RPD is based on the analysis of the frequency of comparisons of atomic situations with indicative features.

If in the result of the processing of open web-resources $N^{(PostPart,ElementIndicator)}$ atomic situations are compared with the indicative characteristic (10). Then the frequency of this indicative characteristic in atomic situations is the following:

$$IndicatorFrequency(ElementIndicator) = \frac{N^{(PostPart,ElementIndicator)}}{N^{(PostPart,Processing)}}, \quad (10)$$

The relevant criterion for the frequency of the indicative characteristic is:

$$IndicatorFrequency(ElementIndicator) \geq \alpha^{(IndicatorFrequency)}, \quad (11)$$

where $\alpha^{(IndicatorFrequency)} \in (0,1]$ - the minimum allowable value for the frequency of appearance of the indicative characteristic.

If the indicative characteristic does not satisfy the criterion (11), it may be due to the presence of gaps in the content of the corresponding RPD element. Gaps on the basis of the frequency analysis of the appearance of indicative features are eliminated by the following methods [3, 19, 20]:

1) use of new open web-resources that are better suited to indicative features that do not meet the criterion (11);

2) reassessment of the conformity measures $\mu(ElementIndicator)_i$ for the indicative characteristic (7);

3) removal of indicative characteristic, if the RPD element is the O, the OA or the classifier term;

4) adding of new markers to marker set $Marker(ElementIndicator)_i$.

The procedures of computer-linguistic formation of documents are developed on the basis of comparing the content of open web-resources with elements by using indicative features, as well as procedures for detecting gaps and incorrect information in documents based on analysis of its structure and content, which allow to improve the quality indicators of documentation and are necessary for the construction of system for the automated formation of documentation [14, 17, 18, 20].

3. Conclusions

Remote-piloted aviation need quality documentation to provide a wide range of services. The development of such documentation will ensure the accelerated development of airspace infrastructure.

The need constantly to improve the quality of remote-piloted aviation documentation explains the importance of developing a method for evaluating its quality based on the needs of the consumer. This is an actual scientific-applied challenge.

Procedures for detecting gaps in documentation have been developed by analyzing the structure of documentation, its content, frequency analysis of markers and indicative features that allow to improve functional quality indicators. This allows to develop procedures for the detection of incorrect information in the RPD, which can improve the quality indices of documentation. Such procedures are very simmilar to the same procedures concerning remote-piloted vehicles.

References

1. Марків О. Формування туристичної документації на основі виявлення та заповнення прогалин, Інформація, комунікація, суспільство (ІКС-2015), Славське-Львів, 230-231 (2015)
2. Пелещишин А., О. Сопрунюк, О. Тимовчак-Максимець: Роль прогалин у документації в процесах виготовлення та експлуатації продукції, Комп'ютерні науки та інформаційні технології, 17-22 (2011)
3. Chaves, Marcirio; Trojahn, Cássia Towards a Multilingual Ontology for Ontology-driven Content Mining in Social Web Sites, Proceedings of the ISWC 2010 Workshops, Volume I, 1st International Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web (2010)
4. Hirakawa, G., Nagatsuji, R., Shibata, Y.: A Collection and delivery method of contents in tourism with location information. Proceedings of 2016 19th International Conference on Network-Based Information Systems, 393-396 (2016).
5. Li, Q. S., Wu , Y. D., Wang, S., Lin, M. S., Feng, X. M., Wang, H. Y.: VisTravel: visualizing tourism network opinion from the user generated content. Journal of Visualization, vol. 19, no. 3, 489-502 (2016).
6. Markiv O. Graphical presentation of tourism documentation structure Міжнародна наукова конференція «Інформація. Комунікація. Суспільство», 216-217 (2016)
7. Markiv V.: Analysis of remote-piloted vehicles use and control system description". , Computer sciences and information technologies, No. 843, 347-351, 2016
8. Markiv V.: Justification of remote-piloted vehicles use and metrology supply improvement. 5th Int. Scientific Conf. ICS-2016, 20–21, 2016
9. Mykyichuk M., Markiv V. Metrology tasks of airphotoshooting by remote-piloted vehicle**,** Вісник "Радіоелектроніка та телекомунікації", В-во НУЛП.,№ 874, 57-61, 2017
10. Mykyichuk M., Markiv V. Peculiarities of fractal analysis of remote-piloted vehicles recognition, VI-a Міжнародна науково-практична конференція "Практичне застосування нелінійних динамічних систем в інфокомунікаціях.,. 20–21, 2017
11. Mykyichuk M., Markiv V. Peculiarities of the radio signals and hindrances in the navigation system of the remote-piloted vehicles, Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska, IAPGOŚ, № 8 (1), 40- 43, 2018
12. Neal, J. D., Sirgy, M. J., Uysal, M.: Measuring the Effect of Tourism Services on Travelers' Quality of Life: Further Validation. Social Indicators Research, Volume 69, Issue 3, 243–277 (2004).
13. Neidhardt, J., Rummele, N., Werthner, H.: Predicting happiness: user interactions and sentiment analysis in an online travel forum. Springer, Berlin, Heidelberg, Information Technology & Tourism, Volume 17, Issue 1, 101–119 (2017).
14. Paramonov, I., Lagutina, K., Mamedov, E., Lagutina, N.: Thesaurus-Based Method of Increasing Text-via-Keyphrase Graph Connectivity During Keyphrase Extraction for e-Tourism Applications. Knowledge Engineering and Semantic Web, Kesw 2016, vol. 649, A. C. N. Ngomo and P. Kremen, Eds., 129-141 (2016).
15. Shreya Upadhyay ; Vishal Pant ; Shivansh Bhasin ; Mahantesh K. Pattanshetti Articulating the construction of a web scraper for massive data extraction 2017, Second International Conference on Electrical, Computer and Communication Technologies (ICECCT) pp.1-4.
16. Tkachenko S. O. Soprunyuk, V. Tkachenko, I. Solomko Efficiency enhancement of optimal reduction method by strengthening parallelism of structural models formation, Machine Dynamics Research, Poland, Vol. 3, No. 2., 85-90 (2013).

17. Vidya Sagar Ponnam ; V. P Krishna Anne ; Venkata Kishore Konki Hidden web data extraction using wordnet ontology's Publication Year: Conference on IT in Business, Industry and Government (CSIBIG), 1-4, (2014)
18. Zhezhnych, P., Markiv, O.: A linguistic method of web-site content comparison with tourism documentation objects. Proc. of 12th Inter. Scientific and Technical Conference Computer Science and Information Technologies (CSIT'2017).
19. Zhezhnych P. O. Soprunyuk Analysis of the tourism documentation quality improvement , Materials of the 7th International Scientific and Technical Conference "Computer Science and Information Technologies (CSIT')", Lviv , 34-36 (2012).
20. Zhezhnych P. O. Soprunyuk Formation of the consolidated data resource of tourism information, VIII Міжнародна науково-технічна конференція "Computer Sciences and Information Technologie, (CSIT'2013), 194-195 (2013).
21. Zhezhnych P., Markiv O. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects. In: Shakhovska N., Stepashko V. (eds) Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, vol 689. Springer, 656-667 (2018)
22. Lytvyn, V., Vysotska, V., Pukach, P., Brodyak, O., Ugryn, D.: Development of a method for determining the keywords in the slavic language texts based on the technology of web mining. In: EasternEuropean Journal of Enterprise Technologies, 2/2(86), 4-12 (2017)
23. Lytvyn, V., Vysotska, V., Pukach, P., Bobyk, I., Uhryn, D.: Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology. In: Eastern-European Journal of Enterprise Technologies, 4/2, 10-18 (2017)
24. Chen, J., Dosyn, D., Lytvyn, V., Sachenko, A.: Smart Data Integration by Goal Driven Ontology Learning. In: Advances in Big Data. Advances in Intelligent Systems and Computing. – Springer 2017. P. 283-292 (2017).
25. Su, J., Vysotska, V., Sachenko, A., Lytvyn, V., Burov, Y.: Information resources processing using linguistic analysis of textual content. In: Intelligent Data Acquisition and Advanced Computing Systems Technology and Applications, 573-578, (2017)
26. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Germany: LAP LAMBERT Academic Publishing (2016).
27. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
28. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
29. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Germany: LAP LAMBERT Academic Publishing (2018)
30. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
31. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
32. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
33. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
34. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of

Ukraine. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
35. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. . In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 100–109 (2017)
36. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, 95–99 (2017)
37. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
38. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
39. Hamon, T.: Biomedical text mining. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf (2017)
40. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 47–55 (2017)
41. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)
42. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)
43. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
44. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
45. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)
46. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st Inter.l Conf. Computational Linguistics and Intelligent Systems, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
47. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf.
48. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
49. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)
50. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
51. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 133–134 (2017)

Method for Automatic Collocation Extraction from Ukrainian Corpora

Maria Kuzmina, Svitlana Petrasova

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

marika958034@gmail.com, svetapetrasova@gmail.com

The article deals with the methods for automatic collocation extraction from Ukrainian corpora. The task of collocation extraction is considered in terms of a corpus-oriented approach [1], based on statistical measures. The term "collocation" is defined as a non-random combination of two words that go together regularly.

Nowadays, the interest in collocation studies relates to the developing of new search engines and machine translation systems, technologies for text recognition, and is due to high frequency of word combinations in text corpora.

The important task of modern corpus linguistics is the extraction of relevant linguistic information, in particular through the use of statistical methods. Additionally, corpus studies allow verifying linguistic theories and hypotheses, as well as identifying and interpreting new language facts [2].

Therefore, the problem of word compatibility or collocation extraction from text data is one of the up-to-date challenges in corpus linguistics.

Linguistic information is extracted and analyzed from the corpus using special programs, namely, corpus managers. The well-known general-purpose corpus managers are SARA, XAIRA (BNC), CQP, which are designed to search data and obtain statistical information from a corpus.

To extract collocations, various association measures can be used. For instance, MI, PMI, t-score, Dice measure, Log-Likelihood are applied to calculate the degree of closeness between components of word combinations in a text corpus. The main drawbacks of statistical methods are noise extraction and ignoring of syntactic correlations between words in long distances.

Analyzing statistical methods for collocation extraction [3], the MI measure has been chosen as a tool for automatic collocation extraction in the Ukrainian-language corpus. MI (Mutual information) measure is used to determine the occurrence of two words by comparing the frequency of their co-occurrence with the product of frequencies of their independent occurrence in the text [4]. One of the advantages of MI is that it allows highlighting key terms that characterize the subject area.

For extracting collocations a corpus of technical instructions has been developed. The corpus meets such requirements as: representativeness, balance, selection, machine readability and standard. The corpus includes three categories (subcorpora) of instructions for existing mobile devices of certain companies. Each text file (instruction) consists of 16 000-17 000 words. The total volume of the developed

corpus is 197 624 words. The proposed algorithm for extraction of collocations from the developed text corpus is as follows:

1. Calculating the total number of word forms in the corpus.
2. Defining absolute frequencies of all the words.
3. Defining the frequency of bigrams.
4. Calculating the closeness value for word pairs using the MI measure. To normalize the values, the MI measure is modified by using the MI3 metric (raise the value to the third power).
5. Calculating dispersion and identifying the threshold for selecting the bigrams: +0.5 from the minimum value.

The designed implementation allocates two-word collocations of several types: terminological and general-language word combinations, proper names, phrases that characterize the topic of the text, as well as some free collocations (fig.1).

```
('датчик', 'наближення'):36193.37
('наближення', 'придбайте'):57365.14
('сторонніх', 'гарнітур'):84038.41
('гарнітур', 'аксесуарів'):36193.37
('аксесуарів', 'металевими'):36193.37
('металевими', 'брелоками'):36193.37
('вплинути', 'прийом'):93558.51
('задня', 'кришка'):36193.37
```

Fig. 1. Implementation of MI3 for Collocation Extraction

The extraction of collocations as statistically significant units allows automating the processing of natural-language information, as well as obtaining data about the mechanisms of phrase formation for their further analysis.

All the mentioned methods do not allow receiving the entire range of collocations from analyzed texts. Thus, there is a need to improve the technologies for collocation extraction from natural-language texts, particularly Ukrainian-language texts. In the future research we intend to broaden the scope of the study on collocation extraction using both statistical and syntactical approaches.

References

1. Khohlova M.V. Experimental verification of methods for collocation extraction // Corpus approaches. – Helsinki. – 2008. – P. 343-357.
2. Bobkova T. Classification of Collocation: The Main Approaches and Criteria // Respectus Philologicus. – No. 29(34). – 2016. – P. 87–98. [in Russian]
3. Petrasova S.V., Khairova N.F. A logical and linguistic model for identification of collocation similarity // Bulletin of NTU “KhPI”. Series: System analysis, control and information technology. – Kharkiv, NTU “KhPI”, 2015. – No. 58 (1167). – P. 14–17.
4. Evert S., Krenn B. Methods for the qualitative evaluation of lexical association measures // Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. – Toulouse, France, 2001. –P. 188–195.

DFA Method for the Analysis of Long-Range Correlations: Application to Statistical Linguistics

L. B. Ivanitskyi and O. S. Kushnir

Optoelectronics and Information Technologies Department,
Ivan Franko National University of Lviv, 107 Tarnavsky Street, 79017 Lviv, Ukraine

lubomyr.ivanitskiy@gmail.com, o.s.kushnir@lnu.edu.ua

Long-range correlations are present in various time series related to many complex systems [1]. To analyze them quantitatively, one can make use of a so-called fluctuation analysis (FA) technique. In brief, it analyzes the mean-square fluctuation $F(w) \sim w^{\alpha}$ as a function of time window size w and finds the exponent α that quantizes presence (or absence) of the long-range correlations in a series, and character of those correlations. In case of non-stationary series with variable statistical moments, i.e. trends available in the series, the FA is insufficient and must be replaced by a detrended fluctuation analysis (DFA) [2, 3]. In spite of a great attention of researchers to this field in the recent decades, we believe that some methodical, technical and even principal moments of the DFA method still need their clarification. In particular, this concerns the case if DFA is applied to such fields as, e.g., a statistical linguistics [4]. Analysis of fluctuations in linguistic systems has a number of peculiarities. In particular, the appropriate time series in many cases include only 1's and 0's, depending on whether the condition of availability of a linguistic element (a given letter, n-gram or word, a word of some length, etc.) at a certain 'time' position in a text is true or false. Moreover, the linguistic time series are 'sparse' in the sense that 1's occur mainly with the relative frequencies as small as $f = 0.01$ or less.

Using Python, we have developed a number of computer programs: (1) an 'extracting' program that assigns a time series to a given symbolic linguistic sequence, (2) a program for generating stochastic (noisy) time series with prescribed exponent values α (where the case $\alpha = ½$ corresponds to a white noise with no long-range correlations), which is based upon a standard Fourier-filtering technique, and (3) a program for analyzing time series and calculating α for the cases of DFA-n of different orders n, with a so-called double passing (see [1]). Continuous time series with the terms varying in the regions [–1; 1] and [0; 1] have been studied, as well as discrete series that involve 0's and 1's or –1's and 1's.

Below we describe in brief the main points investigated by us and the appropriate conclusions.

1. Our analyzing program reproduces with sufficient accuracy the exponents α introduced by the generating program, at least in the region 0.2÷1.2 tested by us.
2. The mean α value for the case of white noise depends very weakly on the series size L in the region 10^{10}÷10^{21}, while the standard deviation $\Delta\alpha$ (i.e., an error of

estimating α) decreases with increasing L according to the power law $\Delta\alpha(L) \sim L^{-a}$, with a being close to the value ½ that follows from the central limit theorem (cf. also with the data [5]). Since we have $\Delta\alpha$ less than 0.01 beginning from $L_{\min} \approx (1\div5)\cdot10^4$, practical difficulties associated with preventing finite-size effects and enabling reliable DFA data can occur for many linguistic systems, which are often shorter than $L_{\min}$.

3. The influence of fitting methods built-in in Python (linear fitting in log-log scale or nonlinear power-law fitting $F(w) = Aw^{\alpha}$ (see [6, 7]) used to derive α with DFA-n (n = 0, 1 and 2) has been studied on a set of 100 binary time sequences (white noise; $L = 5\cdot10^4$; the frequency $f = ½$ for 1's). The data incline to a counter-intuitive conclusion: the fitting method affects insignificantly the results of the scaling analysis and, moreover, the linear fitting yields in somewhat lower square deviations $\Delta\alpha$. Almost the same results are obtained for the case of correlated binary series $\alpha = 0.3\div1.8$. Notice also that our results confirm a known fact that the method DFA-0 is in no case valid for the non-stationary series with $\alpha > 1$.
4. In case of 'sparse' discrete time series with no long-range correlations (the time length $L = 5\cdot10^4$ and the frequency region $f = 10^{-5}\div5\cdot10^{-3}$ for 1's), the nonlinear fitting of the DFA-0,1,2 data becomes superior, because the appropriate α's are closer to the theoretical value ½ for the white noise. Furthermore, these series reveal a crossover phenomenon at $w_{co} = 50\div6000$, with $w_{co} \sim f^{-0.8}$. It is interesting that this crossover has nothing to do with the known competition of trends and fluctuations (see [3]). As a result, one has to reckon with the following problem for the linguistic sequences: since both of the minimal and maximal time windows are limited within the DFA ($w_{\min} > w_{co}$ and $w_{\max} < 0.1L$), there can be a situation when the optimal text-window region $w_{\min} < w < w_{\max}$ for the analysis does not exist at all. The reasons for limitations put on the smallest windows deserve further studies.
5. Even with no accounting for sparseness of the time series and in case of the simplest discrete sequences of 0's and 1's taken with the same frequencies, the FA method appears to be inferior in many respects to any of DFA-n. This implies that the DFA has clear preferences over the FA when being applied to symbolic sequences dealt with in the statistical linguistics.

References

1. J. W. Kantelhard. In: Mathematics of Complexity and Dynamical Systems, Ed. by R. A. Meyers (New York: Springer), 463 (2011).
2. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons and H. E. Stanley. *Nature* **356**, 168 (1992).
3. Kun Hu, P. Ch. Ivanov, Zhi Chen, P. Carpena and H. E. Stanley. *Phys. Rev. E* **64**, 011114.
4. W. Ebeling and A. Neiman. *Physica A* **215**, 233 (1995).
5. P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado and J. L. Oliver. *Phys. Rev. E* **79**, 035102(R) (2009).
6. A. Clauset, C. R. Shalizi and M. E. J. Newman. *SIAM Rev.* **51**, 661 (2009).
7. M. L. Goldstein, S. A. Morris and G. G. Yen. *Eur. Phys. J. B* **41**, 255 (2004).

Distinguishing between Natural and Random Texts: a Statistical Measure Linked to Word Clustering

O. S. Kushnir, A. I. Kashuba and V. V. Yaremkiv

Optoelectronics and Information Technologies Department,
Ivan Franko National University of Lviv, 107 Tarnavsky Street, 79017 Lviv, Ukraine

o.s.kushnir@lnu.edu.ua, andriy.kashuba@lnu.edu.ua

Recognition of natural, i.e. semantically filled messages, as opposed to random or randomized (semantically empty) symbolic sequences, represents an interesting problem. It is sometimes believed that the problem can be solved using the approaches of statistical linguistics (see, e.g., [1–3]).

In this work we try to solve the problem mentioned above, using a suitable circumstance that semantically filled texts reveal a property termed as 'burstiness', 'intermittence' or 'clusterization' [4, 5]. Given the temporal positions t_i ($i = 1, 2, ..., F$, with F being the absolute frequency) of different tokens of some word type w in a text, one can calculate the inter-occurrence, or waiting, times τ_i for this word as $\tau_i = t_{i+1} - t_i - 1$. Although there are more complex and reliable measures of temporal clustering of the word w, one can restrict the analysis to the simplest clustering coefficient $R = \Delta\tau / \bar{\tau}$ [4], where $\bar{\tau}$ denotes the mean waiting time and $\Delta\tau$ the corresponding standard deviation. Due to a stochastic (Poisson-like) nature of the time occurrences t_i, the probability $p(\tau_i)$ is given by a negative exponential distribution, so that we have $R \approx 1$ (or the absence of any 'interactions' of different tokens for a given word type) for a large majority of word types. The exception is so-called keywords, for which $R > 1$ or even $R >> 1$ (a case of clusterization of tokens), and rare and untypical words with $R < 1$ ('repulsion' of tokens).

The main idea of this work is to characterize a text by the mean $\bar{R}$ and the standard deviation $\Delta\bar{R}$ found by averaging over all the word types in this text. They represent cumulative measures of the clusterization effect in the text and, therefore, a peculiar metrics for the presence of keywords and semantics in it. By definition, the word tokens in a random text do not 'interact' with each other, so that we have $R = 1$ for all of the word tokens, except for those rare types which suffer finite-size effects. Then the equality $\bar{R} \approx 1$ immediately follows, whereas $\Delta\bar{R}$ should remain small enough. On the contrary, the $\bar{R}$ parameter for any semantics-bearing text can deviate notably from unity, while $\Delta\bar{R}$ can become relatively large.

To test the above hypothesis, we have analyzed a number of natural and random texts, using an original program written in Python. The natural (fiction) texts have been taken from `http://www.gutenberg.org/`. Among random texts, we have studied the known Miller's monkey texts with different alphabet sizes M, the natural texts randomized (locally or globally) on the linguistic level of words (over 10^9

COLINS'2018, Volume II: Workshop. Lviv, Ukraine, June 25-27, 2018, ISSN 2523-4013
http://colins.in.ua, online

randomization cycles), the texts composed according to a preferential-attachment model by Simon, and the Chomsky texts.

Some measures have been taken in order to reduce the finite-size effects. For this aim, we have neglected the statistics of token occurrences for the word types with the absolute frequencies less than some threshold F_{th} (e.g., $F < 20$) and/or for the types for which the relative frequencies f ($f = F/L$, with L being the total text length in the units of words) are less than a corresponding threshold f_{th} (e.g., $f < 10^{-4}$). Besides of these frequency filters, the statistical significance of rare word types has been reduced by calculating the $\overline{R}$ and $\Delta\overline{R}$ parameters, using weighting coefficients proportional to the absolute frequencies of the words. To further evaluate the negative effect of poor statistics typical for the low-frequency word types, we have compared the data for $\overline{R}$ and $\Delta\overline{R}$ with the results obtained using an improved version of clustering coefficient, which has been introduced in the work [6]. Table 1 exemplifies the data obtained with the only filter with $F_{th} = 20$ and no weighting. Even under such unfavorable statistical conditions, there is a clear distinction of natural and random texts. The $\overline{R}$'s for the latter texts are closer to unity and $\Delta\overline{R}$ notably less than those obtained for the natural texts. In particular, it is instructive to compare $\overline{R}$ and $\Delta\overline{R}$ for the initial natural text and its randomized counterpart.

Table 1. Statistical characteristics of some natural and random texts (see the text).

#	Text	Text length L, 10^5	$\overline{R}$	$\Delta\overline{R}$
1	Natural	1.9	1.33	0.61
2	Randomized natural (10^9 cycles)	1.9	0.97	0.14
3	Monkey ($M = 1$)	10.0	0.96	0.05
4	Simon	1.7	1.02	0.19

Note that the algorithm for producing Simon texts is such that the initial word types from a 'bag of words' are repeated many times in the very beginning of the text and so can, in principle, reveal a spurious clustering in this region, thus increasing the $\overline{R}$ parameter in an uncontrollable manner. However, contrary to these intuitive expectations, it has turned out that the Simon texts reveal the same (nearly unit) $\overline{R}$ coefficients, like the other semantically empty texts.

References

1. A. Cohen, R. N. Mantegna and S. Havlin. *Fractals* **5**, 95 (1997).
2. R. Ferrer i Cancho and R. V. Sole. *Proc. Roy. Soc. Lond. B* **268**, 2261 (2001).
3. D. Kimura and T. Tanaka-Ishii. *J. Natural Language Processing.* **18**, 119 (2011).
4. M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz and A. M. Somoza. *Europhys. Lett.* **57**, 759 (2002).
5. Barabási A.-L. Nature **435**, 207 (2005).
6. P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado and J. L. Oliver. *Phys. Rev. E* **79**, 035102(R) (2009).

Embedding Speech Recognition Tools for Custom Software: Engines Overview

Arthur Dovbysh [1] and Vladyslav Alieksieiev [2]

[1,2] Department of Applied Mathematics, Lviv Polytechnic National University, Bandery str., 12, Lviv, Ukraine, 79013
arthurdovbysh@gmail.com[1], vladyslav.i.alieksieiev@lpnu.ua[2],
https://orcid.org/0000-0003-0712-0120 [2]

Abstract. Different solutions and tools for speech recognition are now available. Nevertheless, implementation of natural language processing still remains a current problem. Developing any custom software with a good style of UI/UX requires the integration of speech recognition. Evidently, the most common solution is to use some engine as an embedded standard tool. Here in the paper we are presenting an overview and an analysis of some popular speech recognition engines: Google Speech Recognition API, Microsoft Speech API, Yandex Speech Kit and Julius. These speech recognition tools are a ready-to-serve and suitable to supplement your own software with a reliable voice command detection or voice control feature. The results of our analysis comes from an experiment of voice recognition using these tools as an embedded component in a custom software.

Keywords: speech recognition, speech engine, API, voice command detection, voice control, Google, Microsoft, Yandex, Julius, overview and analysis

1 Introduction

The modern world is heading for full automation of most processes in your life. One of the main problems is speech recognition for transmitting commands to various devices in spoken form. This may include conversion speech to text or understanding the command directly. However, the recognition accuracy has not yet reached its peak and people try to increase it every day. Therefore, the problem of speech recognition raises nowadays as an extremely serious one and it is crucial for communication of a human and a machine. Managing objects using the language would open wide prospects for automation in many fields of human activity, also would the ability to communicate with machines, especially for users of personal computers who do not know any programming languages. Language contact facilitates the recording of data in the machine and helps humans to work with the computer in a real-time: the man said – the machine has done.

Voice recognition systems are computing systems that can determine a speaker's speech from a common stream. This technology relates to the speech recognition

technology, which converts spoken words into digital text signals by means of speech recognition process by machines. Both of these technologies are in parallel use: on the one hand, to identify the voice of a particular user and, on the other hand, to identify voice commands through speech recognition. Big AAA-companies, like Google, Microsoft, Facebook, Apple, IBM etc., are engaged in the development of such engines and tools. For now, we already have some products based on these systems: Siri, Cortana, Google Assistant, IBM Watson etc. Therefore, speech recognition systems are tools that are ready to use and we can integrate them into our applications.

The main goal of present research is to determine the most reliable audio speech recognition system (speech engine) based on an open source code or a closed source.

2 Some key principles of how speech recognition engines work

When you pronounce a voice request, such as a destination address, then the smartphone does not hear the street and the house number. Your phone defines the sound signal in which the sounds move smoothly one after another, without clear boundaries. The task of the speech recognition system is to restore that signal by such speech of the man. It should be noted that the same phrase is said by different people in different conditions will give various signals. Acoustic modeling system helps to interpret them correctly. After voice request, it is recorded by the smartphone and refers to the server, where the level of interference is determined, is carried out noise cleaning and separation of a useful signal.

Then the record is divided into small fragments – frames. For example, with length of 25 milliseconds and with step of 10 milliseconds. These constants (25 and 10) picked up empirically and have not any math or theoretical proofs. The ends of the frames are superimposed on each other. Thus, out of 1 second of speech, 100 frames are output [1, p. 111–134].

First, each frame is passed through the acoustic model. Then the system of machine learning (with a neural network) presents variants of said words and context. The accuracy of the results directly depends on the completeness of the phonetic alphabet of the system.

For each sound, a complex statistical model, which describes the pronunciation of this sound in the language, is initially constructed.

The system of recognition submits an incoming speech signal to phonemes, and already builds words from them [2, p. 34–58].

For example, the phonetic dictionary of the Yandex system consists of approximately 4,000 elementary units, which include phonemes, their parts and combinations [3–4].

Each frame is compared not with one phoneme, but with a few that fit with different degrees of probability. In addition, the system considers the probability of transitions and determines which frames can go after a specific phoneme. To do this, data on pronunciation, morphology and semantics are determined. In this way, the

system selects word variants. Which then analyzes the forms, parts of the language and possible statistical links between them.

Then the language model is included in the process by which the system determines the most likely order of words and, if necessary, restores unrecognized words for meaning, based on the context and the obtained statistics.

As a result, the received information enters the main unit of the recognition system – the decoder. The software component combines data from acoustic and linguistic models and, based on their association, gives a final result in a form of most probable sequence of words [5].

Thanks to machine learning, the system is resistant to noise and is able to recognize the language with emphasis. Accuracy of modern speech recognition systems exceeds 90%.

3 Speech recognition engines overview

For the research, we have picked four speech recognition systems: Google Speech Recognition API [6–7], Microsoft Speech API [8–9], Yandex Speech Kit [3–4] and Julius [10]. The las one is an open source (Julius) and the others are closed source. It is also very important that Julius are paid, but Google, Microsoft and Yandex provide free SDKs for your use.

The essence of the experiment is to recognize a short poem about life. The text of the poem:

```
Simple Sam was a simple man.
He lived each day by a simple plan.
Enjoy your life and live while you can.
Make each day count and take a stand.
```

In the experiment to determine the quality of speech recognition by these systems was attended by 20 people, including 14 male and 6 female speakers. All speakers were not native English speakers but have good speaking skills (B1+). The average age of participants in the experiment was 22 years. There were no professional speakers in the group. To evaluate the results of automatic recognition, we determined such an indicator as the percentage of correctly recognized words WCR (Word Correctly Recognized) for each speaker, as well as the total final coefficient for each software product.

$$WCR = \frac{H}{T} \bullet 100\%,$$

where H is a number of correctly recognized words and T is total number of words to recognize. The formula is very simple to use.

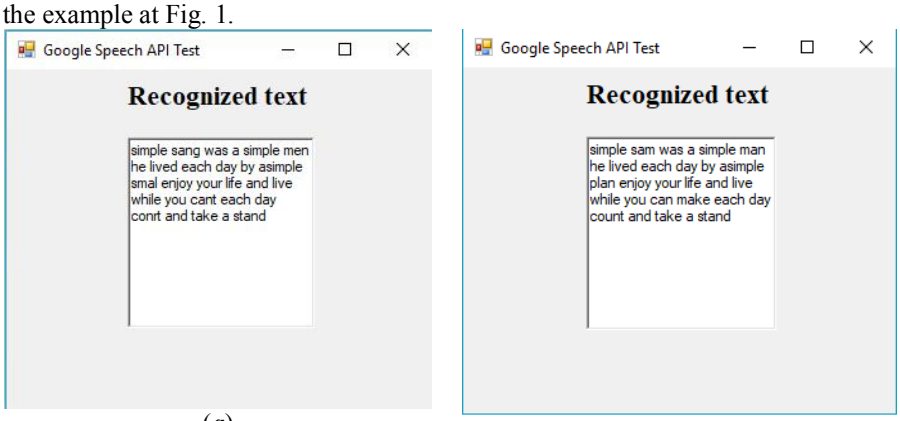
3.1 Google Speech Recognition API

This is a Google product allowing you to enter voice search using speech recognition technology. The technology has been integrated into mobile phones and computers, where you can enter information by voice. Since June 14, 2011, Google announced the integration of the speech engine into Google Search and since then it has been running in stable mode since that time. This system uses the Hidden Markov models algorithm mixed with algorithm of dynamic transformation of time (DTW). Thus, it must works well.

In order to use this Google technology, you need to do the following:

It is necessary to make a POST request to the address (https://github.com/gillesdemey/google-speech-v2) (now it changes frequently – for example, in May there were three changes and therefore it is necessary to be ready for this) with audio data in the FLAC or Speex format. You can implement a demonstration of WAVE-file recognition using C# as we did. The number of restrictions of requests per day did not notice. There was a risk with 10,000 signs, like many other speech recognition systems. We have proved experimentally such values can be overcome daily [6].

Generally, this system was easy to integrate into small WinForms application (see the example at Fig. 1.



(***a***) (***b***)

Fig. 1. Speech recognition with Google Speech API: bad result (***a***) and successful (***b***).

As you can see, it works very well sometimes and sometimes fails. Here are only two of 20 results shown, but WCR for this engine is equal to 81.92%.

3.2 Yandex Speech Kit

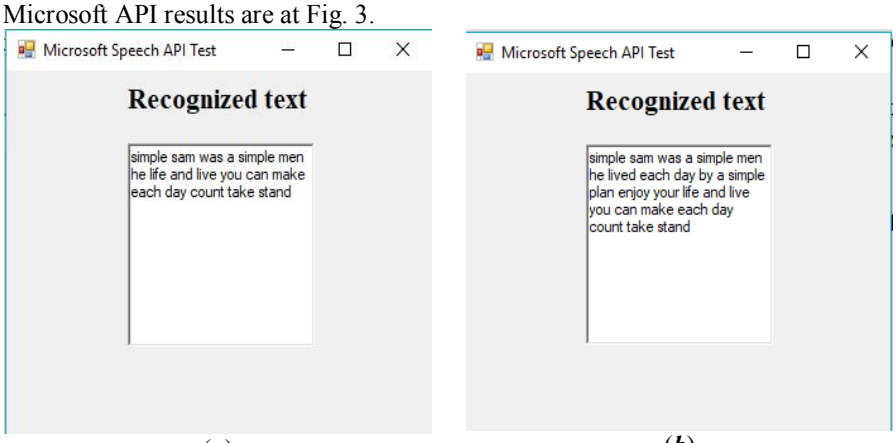
In general, this system works like the previous one. To learn more about you can read the documentations of the product on Yandex APIs web-site. This kit easy to integrate into any applications you want using C#, JS, Go and many other programming languages [3]. The interesting fact is that Yandex's voice search system created based

on this API. Therefore, we tested our poem and most of results were similar to that one at Fig. 2. Finally, this tool have WCR = 64.81%.

Fig. 2. Yandex speech recognition result

3.3 Microsoft Speech API

Microsoft's API [8] has its own SDK (Software Development Kit). This tool is very easy to integrate into your application, in C# particularly. Although, it also has some restrictions. The first one is that we need to configure dictionary of recognized words by the system manually. This dictionary called "Choices". Next, you need to add language settings and create grammar based on Choices and settings [9]. We implemented for test the same application as we did with the Google Speech API and Microsoft API results are at Fig. 3.



(*a*) (*b*)

Fig. 3. Microsoft Speech Recognition API test results

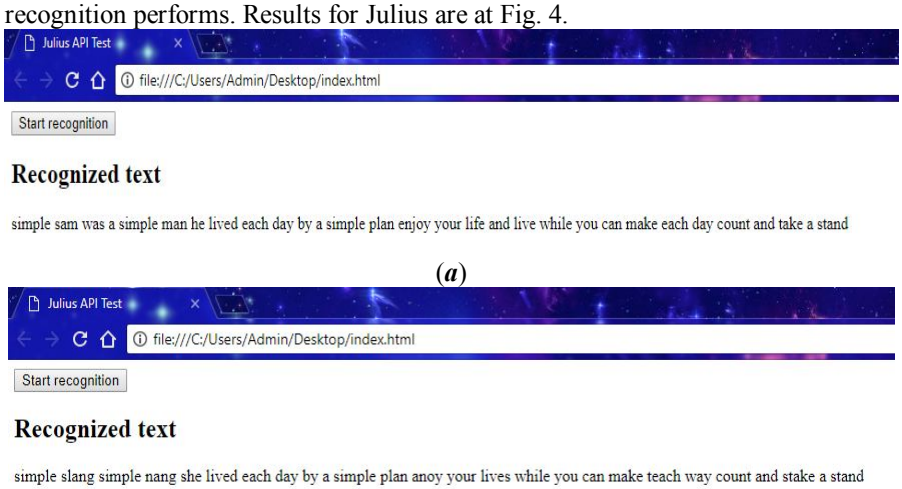
Because of Choices, if system did not recognize words specified in the dictionary – it did not offer any other words. It simply ignore that input signal as you can see on Fig. 3 (*b*). Nevertheless, in other cases WCR of process is quite big. In general, this API get WCR value equal to 79.03%.

3.4 Julius – an open source speech recognition engine

Julius is a high performance continuous speech recognizer with a large vocabulary (large vocabulary continuous speech recognition), a decoder software for research in related speech and development. It is ideal for near-real-time decoding on most existing computers, with a dictionary of 60 thousand words using the trigram word task and a context-independent Hidden Markov model. The main feature of the

project is full embeddability. It is also safe modulation can be independent of model structures and various types of Hidden Markov models, which supports the General state of triphons and the associated mixture-models with a variety of potions, phonemes, and statements. Standard formats are active through free modeling tools. The main platform of the Linux system and other UNIX-like stations, also the system runs on Windows. Julius is open source and distributed with BSD license type [10]. This system originates from Japan [11–12]. The last "officially" mentioned at GitHub scientific publication about Julius was in 2009 [13].

As we have mentioned before this system is paid. However, for test we get free trial for 7 days. With help of Julius SDK for JavaScript we created simple application that records input signal. Then program send it to Julius servers where speech recognition performs. Results for Julius are at Fig. 4.



(***a***)

(***b***)

Fig. 4. Julius API test: successful (***a***) and with some faults (***b***)

Some text was replaced by Julius system (see Fig. 4 (*b*)), because it cannot correctly recognize input speech and engine tries to choose the best variant for replacement. In general, this system has a very big WCR index – 89.67%.

4 Results and discussion

All the results were gathered together to the chart at Fig. 5 and in Table 1. First, we would like to highlight our two "leaders". Microsoft Speech API was the easiest for integration and Julius showed the best WCR value. Outsider among the competitors was Yandex SpeeckKit – it revealed both the lowest WCR value and difficulties of integration. The Google Speech Recognition API takes the middle position due to its good integration (equal to Julius, but not so good as Microsoft Speech API) and good WCR value (not as good as Julius, but higher than Microsoft Speech API).

Table 1. Comparative table of Speech Recognition Engines

	Ease of integration (5 pt max)	Free use	Open Source	WCR value
Microsoft API	5	+	–	79,03%
Google API	4	+	–	81,92%
Yandex Kit	3	+	–	64,81%
Julius	4	–	+	89,67%

Meanwhile, considering other options it is important to mention again that both with a high level of WCR you will have to pay for it, when using Julius. The main disadvantage of Julius may be a lack of documentation. While having a well-documented solution and its own SDK both with clear and easy to use implementations in C# the Microsoft Speech API looks to be ideal as a fast to build software solution.

Google Speech Recognition API does not look much better than Microsoft's API, quite the same, as Microsoft Speech API does not look much worth than Google Speech Recognition API. This means you may use any according to your own preferences to use embedded engine for speech recognition in your own applications.

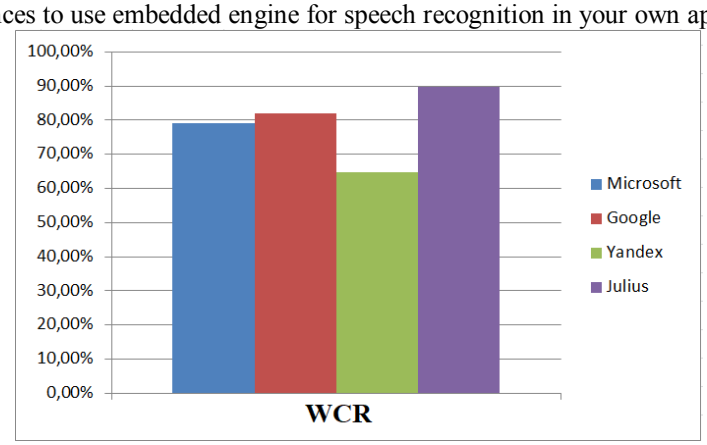


Fig. 5. WCR result chart

5 Conclusions

To summarize, we'd like to highlight Julius showing the best results and this system is easy to integrate. Unfortunately, it is paid. Among the free ones, it is possible to emphasize Microsoft Speech API as well as Google API. These two APIs show good WCR index. Microsoft product is better because it is easier to integrate into your applications. Unlike the Google is not tied to a specific set of words and has its own dictionary. This mean there is a variety to choose.

References

1. Lawrence R. Rabiner and Bernard Gold. Theory and Application of Digital Signal Processing — Prentice Hall, 1975 – 762 p.
2. Plotikov V., Sukhanov V., Jhygulevtsev Yu. Rechevoy dialog v sistemah upravleniya. – Moscow: Mashinostoenie, 1988 – 223 p. – *In Russian* [Речевой диалог в системах управления / В.Н.Плотников, В.А.Суханов, Ю.Н.Жигулевцев. – М.: Машиностроение, 1988. – 223 с. – ISBN 5-217-00148-8]
3. Yandex SpeechKit // Yandex – https://tech.yandex.ru/speechkit/ (Retrieved on May 2018)
4. Yandex.SpeechKit // Wikipedia.org, 18.05.2018 – https://ru.wikipedia.org/wiki/Yandex.SpeechKit (Retrieved on May 2018)
5. Minimum Prediction Residual Principle Applied to Speech Recognition / Itakura F. // IEEE Transactions on Acoustics, Speech, and Signal processing. – February 1975. – Vol. 23, No. 1. – P.67–72.
6. Cloud Speech-to-Text – Speech Recognition // Google Cloud – https://cloud.google.com/speech-to-text/ (Retrieved in May 2018)
7. Google launches an improved speech-to-text service for developers / F. Lardinois // Techcrunch.com – April 9, 2018. – https://techcrunch.com/2018/04/09/google-launches-an-improved-speech-to-text-service-for-developers/
8. Microsoft Speech API // Wikipedia.org, 08.11.2017 – https://en.wikipedia.org/wiki/Microsoft_Speech_API (Retrieved on May 2018)
9. Microsoft Speech Platform SDK 11 Requirements and Installation // Microsoft Developer Network – https://msdn.microsoft.com/ru-ru/library/hh362873(v=office.14).aspx (Retrieved on May, 2018)
10. Julius: Open-Source Large Vocabulary Continuous Speech Recognition Engine / Akinobu Lee // GitHub, April 2018 — https://github.com/julius-speech/julius (Retrieved on May 2018)
11. T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano. "Free software toolkit for Japanese large vocabulary continuous speech recognition." In Proc. Int'l Conf. on Spoken Language Processing (ICSLP), Vol. 4, pp. 476-479, 2000.
12. A. Lee, T. Kawahara and K. Shikano. "Julius – an open source real-time large vocabulary recognition engine." In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1691-1694, 2001.
13. A. Lee and T. Kawahara. "Recent Development of Open-Source Speech Recognition Engine Julius" Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2009.

The Information System for Identification of Content Set Based on Analysis of Similar Texts

Viktoriia Kovalchuk[1], Vasyl Lytvyn[2], Victoria Vysotska[3], Mariya Hrendus[4], Oleh Naum[5]

[1-4]Lviv Polytechnic National University, Lviv, Ukraine
[2]Silesian University of Technologyty, Gliwice, Poland
[5]Drohobych Ivan Franko State Pedagogical University, Drohobych, Ukraine

[1]colourenot@gmail.com,
[2]Vasyl.V.Lytvyn@lpnu.ua,
[3]Victoria.A.Vysotska@lpnu.ua,
[4]mariya.h.hirnyak@lpnu.ua,
[5]oleh.naum@gmail.com

Abstract. Information search is one of the most dynamic areas in terms of development and search for new algorithms for solving this problem. Nowadays a large number of search engines, which work with artistic composition, have been developed, but most of them are oriented only on conducting the work with metadata. In addition, effective methods of analysis of the subject of the text have been used, as well as identification of works after it is considered. Different ways of searching and problems of implementation of an ontologically flexible semantic search are considered. An information system for the identification of a plurality of content has been developed based on a thematic analysis of similar texts.

Keywords: semantic; semantic search; text search; information search; ontology; information; e-library; electronic library; semantic electronic library; text analysis; search.

1 Formulation of the problem

Searching for unstructured information is a labor-intensive process [1-7]. There is an active problem with search in large texts. For the most part, it is considered when working with journalistic works, articles on various sites, etc., but the field of fiction remains aloof. The system that is being developed focuses precisely on searching for artistic texts and should simplify the identification of thematically similar text with plural content that the user enters as a request [8-16].

Typically, apps, currently available on the Internet, are aimed at accurate searches. If a user makes a request that is not in the standard form (for example, an excerpt

from the artistic plan), searching systems for the most part do not handle the search for such scale and nature [17-25].

The impossibility of using as a query a plurality of content, which may comprise, for example, a paragraph of text, prevents the user from searching for texts by comparing fragments of one of them with others in searching of similar [27-35].

The developed system takes into account these nuances and gives the opportunity for user to promptly obtain the necessary information, i. e. the ordered data of the artistic works that correspond to the thematic query [36-41].

2 Information search

Information search is a process of searching for unstructured document information that satisfies the information needs.

Request is a condition for which the search is done. The object of a request is the information unit stored in the searching system database.

The search process consists of a number of operations aimed at gathering information, processing it and giving out the results. The main task of information retrieval is to satisfy the information needs of the user.

Search methods are divided into the following categories:

1. Address search is a search process that is implemented using the formal features specified in the request.
2. Document searchis a search process that occurs in the system of the repository of primary documents, or a database with secondary documents.
3. Factor search is the process of finding the facts which correspond the information request.
4. Semantic search is the process of finding documents by content.

Since the research is based on working with a database of texts, it is worth to admit the features of search in electronic libraries. If we are talking about content search under metadata search, then for electronic libraries it is divided into two types: full text and inaccurate.

The problem of full text search is the absence of understanding of the meaning of the information needs of users, the meaning of sought-after text documents. To solve them, it is necessary to process the documents and formalize the semantics contained in them, using personal knowledge representation models (such as taxonomies, thesauri, and ontologies).

Inaccurate search is widely used in the study of little-known, specific works and works in a foreign language which correct writing of its names is unknown. Using inaccurate search, the user needs to insert into the search bar all the variants of writing the searched word (plural/singular, as well as spelling mistakes) [1].

3 Ontological analysis of the text

To work with the ontological analysis of the text, widely use the method of semantic proximity (semantic similarity). Semantic proximity can be determined between the various components of the triplets. At the same time, as a basic proximity, we can consider the proximity between elements of ontologies (classes, predicates, terms). Classes and predicates are elements of the ontology scheme, and terms are elements of the ontology of subject knowledge in the form of a taxonomy or thesaurus presented using the standard SKOS (Simple Knowledge Organization System) [2].

On the basis of assessments of the semantic proximity between the ontology elements, the proximity between the metadata of the resources of the electronic library can also be determined.

There are a number of effective methods of searching the semantic electronic library:

1. Simple search. Search for information objects based on the use of their lexical tags.
2. Search by graph. This kind of search is close to a simple search but uses the SPARQL query to solve a task. The difference is that the request for an identifier (URI) of this object will be known.
3. Contextual search. Contextual searches for objects are performed by execution a search query for contextual metadata. An example of such a search can be the search for projects, documents and other information objects that include a given set of relationships.
4. Content search. The content search is used for objects that are annotated with sets of triplets of thematic metadata [3].

4 Semantic search

There are different approaches to organizing semantic text searches. In recent years, semantic annotation of the text has become the most popular one. There are various ways to solve the problem of semantic annotation. Each document or part of the document attributes a set of semantically close tags. In the future, you can search documents for these tags. In addition, you can search for documents using regular full-text searches, and then take these tags into account when working with a document, getting more information through them. Typically, people, places, organizations, or other subjects are used as tags [4].

Analyzing all these data, it becomes obvious that one of the most advanced methods of searching for texts is a semantic search. But in its implementation there are a number of problems that impose certain restrictions on its application on small-scale systems (for example, nowadays semantic search operates correctly only using unnatural languages; i.e., the user needs to be familiar with the formal languages to formulate the query).

Methods for automating the process of converting free-form queries (for example, in the form of sentences in the natural language or as plural of the keyword list) to the formal form are currently the subject of the study. The construction of the mapping of ontologies of subject domains to formal queries is also actively explored [5-6].

One way to implement a semantic search is to organize using a method of tagging. It can be done by highlighting the tags from the text books [7].

To match the content of the text, the tag must contain its keywords. A tag can be a keyword, sentence, or a paragraph in which a keyword has encountered. Thus, the task of tagging can be reduced to the search for keywords in the text.

On the other hand, if you use keywords as a signs to split text, you can significantly increase the speed and quality of the splitting by reducing the space of attributes and noise filtering.

The highlighting of key words are divided into two stages. The first stage is the selection of candidates for the keywords. At this stage, stop-words are deleted, parts of the language can be filtered (prepositions and conjunctions, etc.). The second stage is to test candidates for semantic proximity to this text. This stage is productive when analyzing large-scale text, because it allocates the logic of communication of different words, which is difficult to trace without context (which is hard to catch if the keywords are highlighted at different ends of the text). At the same time, there is also the problem of an expert assessment of the proximity of words and the formation of a morphological basis.

5 Conclusions

Based on the collected data and the ways to solve the problem of finding unstructured information, we can conclude that the most effective way for a full text search is a semantic search. It allows you to give the most accurate results, since it handles all text, not just metadata. This is a fundamentally important position, since the information system being developed looks like a kind of electronic library, and the job is to identify the query with a large array of text.

References

1. Alahmari, F., Magee, L.: Linked Data and Entity Search: A Brief History and Some Ways Ahead. Proceedings of the 3rd Australasian Web Conference (2015)
2. Bontcheva, K., Tablan, V., Cunningham, H.: Semantic search over documents and ontologies. Bridging Between Information Retrieval and Databases, Springer (2014)
3. Taylor, William P.: A comparative study on ontology generation and text clustering using VSM, LSI, and document ontology models. Clemson University (2007)
4. Giannopoulos, G.: A tool for semantic annotation and search. In: The Semantic Web: Research and Applications, Springer Berlin Heidelberg, 376-380 (2010)
5. Berlanga, R., Nebot ,V., Pérez, M.: Tailored semantic annotation for semantic search .Web Semantics: Science, Services and Agents on the World Wide Web, 69-81 (2015)

6. Saša, Neši: Semantic Document Architecture for Desktop Data Integration and Management: Doctoral Dissertation (2010)
7. Hotho, A., Staab, S., Stumme, G.: Wordnet improves Text Document Clustering. In: Proc. of the SIGIR 2003 Semantic Web Workshop, 541–544 (2003)
8. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Germany: LAP LAMBERT Academic Publishing (2016).
9. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
10. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
11. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
12. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
13. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
14. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
15. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
16. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter.l Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
17. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. . In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 100–109 (2017)
18. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)
19. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
20. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
21. Hamon, T.: Biomedical text mining. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf (2017)
22. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 47–55 (2017)
23. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)

24. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)
25. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
26. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
27. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)
28. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
29. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf. (2017)
30. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
31. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)
32. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
33. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, 133–134
34. Kirkin, S., Melnyk, K.: Intelligent data processing in creating targeted advertising. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 131–132 (2017)
35. Hordienko, H., Ilchenko, M.: Development and computerization of an English term system in the fields of drilling and drilling rigs. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 129–130 (2017)
36. Gorbachov, V., Cherednichenko, O.: Improving communication in enterprise solutions: challenges and opportunities. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 127–128 (2017)
37. Didusov, V., Kochueva, Z.: Statistical methods usage of descriptive statistics in corpus linguistic. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 125–126 (2017)
38. Verbinenko, Yu.: Discursive units in scientific texts. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 120–123 (2017)
39. Titova, V., Gnatchuk, I.: Evaluation ofa formalized model for classification of emergency situations. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 110–119 (2017)
40. Kuprianov, Ye.: Semantic state superpositions and their treatment in virtual lexicographic laboratory for spanish language dictionary. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 37–46 (2017)
41. Kotov, M.: NLP resources for a rare language morphological analyzer: danish case. In: 1st Inter, Conf, Computational Linguistics and Intelligent Systems, COLINS, 31–36 (2017)

Section II.
Intelligent Systems

COLINS'2018, Volume II: Workshop. Lviv, Ukraine, June 25-27, 2018, ISSN 2523-4013
http://colins.in.ua, online

Reverse-search System of Similar or Identical Images

Oleh Veres[1], Yaroslav Kis[2], Vladyslav Kugivchak[3], Igor Rishniak[4]

Information Systems and Network Department, Lviv Polytechnic National University,
Bandery str., 12, Lviv, Ukraine, 79013
Oleh.M.Veres@lpnu.ua[1], Yaroslav.P.Kis@lpnu.ua[2],
Vladyslav.Kuhivchak.MKN.2017@lpnu.ua[3], rishnyakiv@gmail.com[4]

Abstract. The article algorithms of image analysis on the number of errors of operation are investigated. The prototype of the system was created, and testing of the described methods was carried out. The result of the analysis became the basis for the information system project of reverse search of similar or identical images.

Keywords*:* analysis, detector, descriptor, image, key point, method, pixel, hashing.

1 Introduction

Graphic images are an important part of the information resource. Some types of images are subject to copyright and are protected by law. Therefore, it is necessary to determine exactly which graphic elements are considered similar. Obviously, these are full duplicates that can also be modified.

If, for one author, the illustration in the work under study is a photo of a known painting, then another author can get almost identical image by making a picture of himself. A similar situation with images that are freely accessible and can be used without any restrictions.

The computer program is not able to evaluate the content of the image and make a conclusion about the license, so the final decision is taken by the expert.

2 The Analysis of Recent Researches and Publications

Similar file search methods, can be used MD5 signatures search for completely identical files are locally sensitive hashing to find identical images. However, these methods are not feasible, since even changing the file format (for example, converting JPEG to PNG) completely changes the binary structure of the file. Therefore, the search methods based on image processing as binary files are inappropriate, unless absolutely identical ones are to be found. But in the case of using an image file hashing using the MD5 (or similar function) algorithm, even editing EXIF

information will result in files being considered differently. Therefore, it is necessary to use algorithms that consider the image as a graphic object, and not as a binary file. There are many types and image formats, but they can all be rendered in raster graphics and saved in one of the popular formats (PNG, TIFF, JPEG). Each image (in this case it is a raster graph) consists of individual points - pixels. Each pixel has its own color and position in the image. The most commonly used RGB (red, green, blue) model is the color scheme, where each color is created by combining three basic colors in different proportions. In this way, the color numerical values that are easy to manipulate can be specified. By accepting the maximum depth of 24 bits (~ 16 million colors), there will be the color values from RGB (0,0,0) to RGB (255,255,255), that is, one color may have an intensity from 0 to 255 units. This color descriptor model greatly simplifies image manipulation techniques.

There are two main areas for processing of graphic information: the definition of key points on the image and the use of locally sensitive hashing. These methods can be combined. They give good results in finding similar images. Using Hamming's measure [1], you can find the same type of image, even with 90% cropping the image. The method has a high probability of false results.

Own way of identifying key features developed image A. O. Biloshchitsky and O. V. Dichtyarenko [2]. The image is described using vectors. The method is named min-Hash and tf-idfWeighting.

The most popular search duplicate images are three indexing methods: Average Hash, Difference Hash, Perceptual Hash.

To find similar images, a method is used to select key points [3-4].

So, to find snippets of an image or similar content of the illustrations - you need to experiment with the methods of determining key points, each of which also has its own set of advantages and disadvantages.

The main methods used to construct detectors and descriptors are: FAST [5]; SIFT [6]; ORB [7, 8]; AKAZE [9]; BRIEF [10]; BRISK [11].

Today there are many systems for image recognition. The most popular ones are: TinEye, Google Similar Images, Yandex. Maps, AntiDupl.NET.

TinEye. This is the first mechanism for searching images on the Internet, not using key phrases or metadata, but according to a copy of the image. When downloading an image, this program creates a "unique and compact digital signature or imprint" and compares it with other indexed images in its own database of illustrations. This procedure allows recognizing even strongly altered versions of the original image, but usually does not return similar images in the results. Disadvantages: the service only works with file formats: JPG, PNG and GIF; image size not less than 100x100 pixels; file size is no more than 1 MB; Can not upload image gallery; the impossibility of creating one's own database (DB) for work.

Google Similar Images. Google uses the so-called "Reverse image search" mechanism, which eliminates the need to enter keywords and terms in the Google search field. Unlike TinEye, results may include similar images, web results, image pages, and various image permissions. Disadvantages: the impossibility of loading an

image gallery; Only file formats are supported: JPEG, GIF, PNG, BMP, TIFF or WebP; impossibility to create an own database; no flexible user setup for search criteria; it is impossible to perform a search on other search engines.

Yandex. Pictures. When searching for images using the appropriate section of the Yandex, a list of images which are similar to the one selected can be obtained. Duplicates found are not displayed in search results: the presence of duplicates can be seen on the preview page of the duplicate image. Disadvantages: It does not allow creating one's own image database for the search; cannot search in other search engines; is prohibited to use on the territory of Ukraine.

AntiDupl.NET. This program searches for the same and damaged images on a disk. As a rule, modern computer users have numerous image galleries in different formats, and not everyone wants to keep the same one, while taking the disk space. In order not to perform manual searches, the program is created which automates these actions. The search is based on a comparison of the contents of the file, so it is possible to search not only identical, but also modified (similar) images. Disadvantages: Does not allow creating separate sections in the database of images; with a large amount of data in the database (> 10,000) works extremely slowly or generally generates errors; the database is taken from the current computer; accordingly, all users have free access to it; looks for similar images, but does not structure them for the most similar, therefore it causes additional volume work for the user; the program does not process an image with defects (although the description says that it can process it); the program settings are not stored in the future; the program settings are not stored in the future; when the name of a picture in a database is changed, it issues an error and cannot process it.

The information system must not only find all explicit duplicates (those that have changed only the colors, sizes or format), but also "similar" images, while minimizing the amount of work for the system operator. Also, the system should find images that have undergone modifications that are easy to perform: rotation, reflection, color change, image cropping.

Today there is no information system of image analysis, which identifies the same or similar images from a database created images. The reverse search image information system must compare objects with identical or similar objects in the database to find information about the owners of the illustrations. You must also find modified images. This will prevent the use of plagiarism.

3 Research of algorithms for revealing the similarity of different images

To develop a reverse search information system project, a threshold function was searched for duplicate searches, using hashing on average and Hemming's measure. The threshold function should be chosen between 68-88%. The ORB method was

selected based on the analysis of test methods for constructing detectors and descriptors.

It is necessary to analyze the effectiveness of methods for images that have undergone modifications. The result of the research is the basis for the design of an information reverse image search system.

To conduct research, test modules for the program realization of the prototype of the information system of reverse pattern search was created based on their own data sample [12-18].

The algorithms to identify the similarity of different images will be selected, as well as check for errors in the work of each of the methods. To do this, three groups of images are created: identical images, similar images and different images. Each group contains two tests to verify the validity of each algorithm.

Same pictures. Test 1 – two completely identical images (Fig. 1, a). To determine the similarity of images as resources, two completely identical images were taken, each with key points defined and a descriptor created to remove these points and compare them to identity.

Test 2 – to determine the similarity of images as a resource, two completely identical images were taken (Fig. 1, *a* and *b*), one of which was rotated 90° (Fig. 1, *b*), after which for each of them the key points were determined and a descriptor is created to remove these points and compare them to each other for identity.

Fig. 1. Same pictures: *a*) original; *b*) reversed by 90 degrees; *c*) changed perspective; *d*) turned 45 degrees

Similar images. Test 1 – two similar images (Fig. 1, *a* and *c*), one of which was photographed from a different angle (the main object of the study - the book, is closer) (Fig. 1, *c*). Test 2 – two similar images, one of which was photographed from another angle (Fig. 1, *c*) and the main object of the study is at an angle of 45 degrees (Fig. 1, *d*).

Different images. Test 1 – two completely different images, one of which is an illustration of the radio site “Radiy”, and the other – the main page site of the University of Lviv Polytechnic. Test 2 – two completely different images, one of which is a photograph of the National University "Lviv Polytechnic" (Fig. 2, *a*), and the other - "Ivan Franko Lviv National University" (Fig. 2, *b*).

a)  *b*) 

Fig. 2. Different images for the Test 2: *a*) National University "Lviv Polytechnic"; *b*) Ivan Franko Lviv National University

For each image, key points are defined and a descriptor is created to remove these points and compare them to identity.

The results of the tests carried out are presented in the table 1-4. The basic parameters of test results:

- *PK* – number of key points;
- PK_S – a number of similar key points between two images;
- K_S – the percentage of shared key points is the similarity between two images.

Table 1. Results of the study by the ORB method

Key points	*Same pictures*				*Similar images*				*Different images*			
	Test 1		**Test 2**		**Test 1**		**Test 2**		**Test 1**		**Test 2**	
	Fig1	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**
PK	416	416	416	415	416	419	416	396	423	416	455	454
PK_S	416		406		167		73		27		27	
K_S(%)	100		98		40		19		6		6	

The ORB method is well suited to all tests, since the percentage of shared key points decrease according to less similar images. The number of key points obtained

is practically equal in each of the experiments, taking this into consideration, it can be said that this method proves to be stable.

In the percentage ratio, the AKAZE method shows the results at the level with the ORB, but the number of generated key points here is much smaller and not even, so it can be said that the method is stable in the results, but unpredictable as to the number of creating the main points in the image.

Table 2. Results of the study by the AKAZE method

Key points	Same pictures				Similar images				Different images			
	Test 1		**Test 2**		**Test 1**		**Test 2**		**Test 1**		**Test 2**	
	Fig1	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**
PK	33	33	33	33	33	36	33	35	114	129	326	376
PK_S	33		32		13		8		7		16	
K_S(%)	100		97		36		23		5		4	

Table 3. Results of the study by the BRISK method

Key points	Same pictures				Similar images				Different images			
	Test 1		**Test 2**		**Test 1**		**Test 2**		**Test 1**		**Test 2**	
	Fig1	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**
PK	363	363	363	373	363	389	363	492	369	1009	1330	1444
PK_S	361		333		160		72		11		26	
K_S(%)	99		89		41		15		1		2	

The results of the tests with the use of the BRISK method indicate that this algorithm also coped with its task, but in relation to previous methods, it showed the worst result in finding similar and identical images, but was able to clearly distinguish between different illustrations in the tests. The number of key points is not stable and increases with increasing detail in the image.

Table 4. Results of the study by the FAST method

Key points	Same pictures				Similar images				Different images			
	Test 1		**Test 2**		**Test 1**		**Test 2**		**Test 1**		**Test 2**	
	Fig1	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**	**Fig1**	**Fig2**
PK	566	566	566	566	566	560	566	658	411	742	1777	1863
PK_S	566		23		282		33		27		39	
K_S(%)	100		4		50		5		4		2	

The FAST method was the leader in determining the speed of finding key points and deducting the descriptors for them, but did not cope with the tests, and although the number of its key points was much larger than its predecessors, it did not allow them to recognize identical images when they rotated 90 degrees and similar images when rotated 45 degrees.

4 Structural elements of the system reverse-search image

The analysis of the conducted research allows formulating requirements and executing the design of information system of reverse image search.

To find the same images, we use the average value of the hash method. To deduct a measure of similarity is the Hamming distance. The user, at the beginning of the work, downloads the image from the external media, after which the system reads it and reduces to 100 by 100 pixels. Next, the number of colors is reduced, turning it into a black and white image with shades of gray. After receiving a new image, the average pixel color for it is looked for. To do this, it is necessary to go through all the pixels, adding all their colors individually for each RGB and dividing them by their number. Knowing the average color of the image, we pass through each pixel of the image is passed through, comparing it to the average. A signature in the image: for each step is created, if the pixel is darker than the average, a single signature is added and the pixel value to the black color is assigned, if the light is 0 and white. So a bitmap image that can also be displayed as a binary number is assigned, and a completely black and white image without shades of gray is obtained. The same is done with all images of our own database. All the necessary data from using Hamming distance will be obtained. The Hamming distance is defined as the number of bits that differ between the two corresponding input vectors of a fixed length. The larger this distance, the more different the image. If the Hamming distance is zero, this means that the images being studied are completely identical.

The detector is used to determine the key points used. To search for similar key points there is a descriptor. For this, the ORB algorithm is used. To find the similarity between images, that is, the definition of key points and their comparison, the OpenCV open access library is used.

The design of the system is accomplished by means of structural modeling. For better understanding of the relationship between the system and the external environment, data flow diagrams [12] are constructed. The software component of the system is implemented by means of the programming language [13, 14]: Java - as the main language for the development of business logic and user interface, as well as SQL-for work with the database. The analysis of existing technologies of work with information resource is conducted [15-17]. MySQL is selected to create and work with the database of user images.

The developed information system provides the user with the opportunity to make a selection of images based on the input data, to perform their revision, to find the same and similar images, to add new images to the database, to view the data of the owner of the illustration in order to determine the plagiarism [18-49].

5 Conclusion

The algorithms of image analysis on the number of errors of operation are investigated for the development of the information system project. For this purpose, groups of identical images, similar images and totally different images were created. The FAST algorithm did not cope with this task, and therefore, despite its best results in image processing, this method cannot be used. ORB and AKAZE algorithms showed the best test results for all indicators.

According to the results of the tests, ORB method is chosen for implementation in the reverse image search information system, since it generates much more key point per unit time than the AKAZE method.

The chosen method is implemented in the prototype of the reverse-search information system. The system is designed to compare objects with objects in the database that are identical or similar.

Further work will be devoted to the improvement of the prototype of the information system software and the development of an intelligent component for efficient image searching.

References

1. Shapiro, L., Stockman, G.: Computer vision. Washington University (2006)
2. Biloshchyts′kyy, A., Dikhtyarenko, O.: The effectiveness of methods for finding matches in texts. In: Managing the development of complex systems, 14, 144–147 (2013)
3. Shozda, N.: Searching for textures in large databases. In: Informatics, Cybernetics and Computing. Donetsk. Ukraine, 39, 182 - 187 (2002)
4. Biloshchyts′kyy, A., Dikhtyarenko, O.: Optimize the match search system by using algorithms for locally sensitive hashing of text data sets. In: Managing the development of complex systems, 19, 113-117 (2014)
5. Alcantarilla, P., Nuevo, J., Bartoli, A.: Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. British Machine Vision Conference (BMVC), (2013)
6. Grewenig, S., Weickert, J., Schroers, C., Bruhn, A.: Cyclic Schemes for PDEBased Image Analysis. In: International Journal of Computer Vision (2013)
7. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF, Computer Vision. (ICCV). In: IEEE International Conference, 2564-2571 (2011)
8. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: 9th European Conference on Computer Vision (ECCV), 430-443 (2006)
9. Yang, X., Cheng, K.: LDB: An ultra-fast feature for scalable augmented reality. In: IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR), 49-57 (2012)
10. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary Robust Independent Elementary Features. In: 11th European Conference on Computer Vision, 778-792 (2010)
11. Leutenegger, S., Chli, M., Siegwart, R.: BRISK: Binary Robust Invariant Scalable Keypoints. Zurich, 2548-2555 (2011)
12. Litvin, V., Shakhovska, N. Designing Information Systems: Teaching. Lviv (2000)
13. Mashnin, T.: JavaFX 2.0. Development of RIA applications. BHV-Petersburg. (2012)

14. Davis, M., Phillips, J.: Studying PHP and MySQL. M .: Symbol-Plus (2008)
15. Shakhovska, N., Bolubash, Y., Veres, O.: Big Data Federated Repository Model. In: The Experience of Designing and Application of CAD Systems in Microelectronics. (2015)
16. Shakhovska, N., Veres, O., Bolubash, Y., Bychkovska-Lipinska, L.: Data space architecture for Big Data managering. In: Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), 184-187 (2015)
17. Veres, O.: Shakhovska, N. Elements of the formal model big date. XI International Conference on Perspective Technologies and Methods in MEMS Design, 81-83 (2015).
18. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Germany: LAP LAMBERT Academic Publishing (2016).
19. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
20. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
21. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
22. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
23. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
24. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
25. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
26. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
27. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 100–109 (2017)
28. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)
29. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
30. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
31. Hamon, T.: Biomedical text mining. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf (2017)
32. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 47–55 (2017)

33. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)
34. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)
35. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
36. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
37. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)
38. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st Inter.l Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
39. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf. (2017)
40. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
41. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)
42. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
43. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 133–134 (2017)
44. Kirkin, S., Melnyk, K.: Intelligent data processing in creating targeted advertising. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 131–132 (2017)
45. Hordienko, H., Ilchenko, M.: Development and computerization of an English term system in the fields of drilling and drilling rigs. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 129–130 (2017)
46. Gorbachov, V., Cherednichenko, O.: Improving communication in enterprise solutions: challenges and opportunities. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 127–128 (2017)
47. Didusov, V., Kochueva, Z.: Statistical methods usage of descriptive statistics in corpus linguistic. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 125–126 (2017)
48. Verbinenko, Yu.: Discursive units in scientific texts. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 120–123 (2017)
49. Titova, V., Gnatchuk, I.: Evaluation ofa formalized model for classification of emergency situations. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 110–119 (2017)

Extracting and Classification the Semi-Structured Data of Web-Systems

Irina Pelekh

Lviv Polytechnic National University, Lviv, Ukraine

presty@i.ua

Abstract. The extracting and classification of semi-structured data of web-systems is described. The definition of semi-structured data is given and the main characteristics are defined. The variety of tasks text information processing is grouped into the eleven large classes related to the analysis of text data. The traditional models of knowledge representation are considered. An algorithm for the web-sources, from which data will to be obtained, ontological model integrating creating is proposed. The process of data extracting using the query language to the markup language elements is characterized.

Keywords: semi-structured data, web-system, extracting, classification, ontology.

1 Problem formulation

In our time of global computerization, the information search using the Internet has long been extremely popular and most used. In the era of the Internet the systems of an entirely new type - services began to appear. These are standalone programs that implement certain functionality that developers can use in their applications. The service itself became a separate unit, separated from the user programs internal structure. The service can be written using another programming language, can be operated under the control of another OS and the server on which it operates, can be physically located anywhere in the world. A service whose requests are generated and transmitted over a local area network or the Internet is called a web service. Especially popular, today, are web-services that integrate the data from several information resources.

In WWW, information is provided to the user in the form of web pages that do not have a clearly defined structure. There is an urgent problem of obtaining data from such sources for further work with them. The application of web-scraping method generates the problem of analysis and identification of semi-structured data. Semi-structured data is characterized by the lack of strict table structures and relationships in relational database models, however, this data form contains tags and other markers

for the separating of semantic elements, as well as to provide a hierarchical structure of records and fields in data sets [6].

2 State of arts

The Internet is the largest source of data, most of which are presented as web-pages that do not have a strictly formalized structure. To make a quality search you need to use the complex mathematical models, semantic analysis and other methods of information analysis. Therefore, the data, that is receiving should be structured in a certain way. However, most web-sources databases are presented in the form of unique structures, which makes it difficult to obtain structured data from similarly semi-structured web-pages.
Extracting and classification of structured data from web-pages is reduced to the following tasks [8]:

- searching and receiving of available information resources for data extracting (navigation problem);
- recognition of areas containing the required data (data recognition problem);
- search of the found data structure (the problem of finding a common data structure);
- ensuring of the extracted data homogeneity (the problem of matching the attributes of the extracted data);
- data integration from different sources (problem of data integration).

Traditionally, in the systems of text analyses for knowledge representing are used four types models: productive, formal-logical, framing and semantic-network model. Based on these models the solutions are described and the main prospects for their using [1].

It should be noted the prospects of using functional control systems to solve the problems of text analysis. In this case, the core of such systems may be became the intelligent information systems that include elements of artificial intelligence, based on the methods and means of the intelligence theory [2, 9-12]. In [3] authors propose a system for obtaining data from the science-metric databases that uses the language of queries for markup language elements.

The following basic methods are used to solve the problem of text information processing [2, 7, 8, 13-18]: data mining; associative rules; production model; formal-logical model; framed model; semantic-network model; decision tree; clusters; mathematical functions; etc. [19-49]

The analysis of this methods shows that each of them has a well-developed formal system that allows you to make a sufficiently full description of the entities and processes of various subject areas. However, they all have a significant drawback in their usage. They do not allow you to describe the illogical, incomplete or contradictory of text, which is a reflection of the natural language. There is a paradox

associated with the limited capabilities of logical formal systems and the needing to describe logical features and not logical knowledge and data contained in the texts.

3 Algorithm of extracting and classification the semi-structured sources data

The quality of extracting and classification the necessary information from semi-structured sources depends on the methods and technologies, which are working with such data. So the most effective to solve this problem, in our time, are considered semantically oriented technologies [4-6].

To classify semi-structured data from web-sources, we offer an algorithm for creating an integration ontological model of all web sources from which data will to be extracted.

To classify semi-structured data from web-sources, we offer an algorithm for creating an integration ontological model of all web sources from which data will to be extracted.

Consequently, in the first stage, the writing procedure of the semi-structured html format information into a pre-created database is performed.

The process of extracting data from the HTML page is as follows. Web page returned by the server is formatted using the markup language (mostly HTML), for further displaying in one form or another using a special program (web browser). In fig. 1 is shown an example of a web browser's visualization of a particular data area (social network community) and the source code of this data.

Here, for example, the data description on the community wall is as follows:

```
…
<div class="post_content">
<div class="post_info">
<div class="wall_text"> …
```

A certain number of different classes have been allocated that responsible for their data area. So, the class "wall_text" describes the record recording on the wall. The text importance in the record is determined by the presence of such characteristics: bold type, italic, underline, hashtag, etc.

To get such data, is made the searching for the classes names that are responsible for describing the necessary information and obtaining their content. Thus, one of the database table results fields is filled in. To automate this process, the data extraction program uses the query language for the markup language elements (Xpath).

The next stage is the direct construction an integration ontological model of all web sources from which data will to be extracted, on the basis of the received and created databases. When constructing any algorithm, the primary task is to determine the input and output data. The input data for the algorithm for constructing an integration ontological model of all web sources from which data will to be extracted are:

- structural schemes of the web-sources databases;

— domain ontology.

Fig.1. The data of records description on "Interior and Decor" community wall of soc. network

The domain ontology is usually developed pre-emptively, with the participation of a domain expert and a knowledge expert specialist in ontological format. The process of creating such a model takes a long time, but it only needs to be done at the initial stage of integration. With the further addition of new systems are working in this field, the ontology itself does not require any additional changes.

Output data is a general ontological model that describes the structure of all web-sources within their domain and the relationship between elements of different systems. Such model will be modeled using RDF language, its RDFs extension and OWL language.

The algorithm for constructing an integration ontological model of all web sources from which data will to be extracted contains 6 main steps, namely:

- Representation of the database structure in the RDF form (sequential displaying of the scheme S in the RDF format.
- Adding semantic properties and creating the ontology. This step is realized by using the procedure of determining the database elements common features and adding links between them.
- Adding ontologies of the upper levels and domain ontology. We implement this step by using the OWL language by using the owl: import command. By usage the transitivity rule in RDF, additional ontologies are extending the subject areas and add new concepts and properties.
- Checking of the created ontology. This step is implemented by checking and analyzing the elongated ontology for "connectivity", that is, we check whether there is a lack of semantic links anywhere. If so, then go to step five, if not - go to step six.
- Editing an extended ontology by usage the ontology editor and adding links between concepts. Next, go back to step 4.

- Storing the received total ontology of the system structure in the RDF metadata repository.

4 Conclusion

The need to develop methods and tools for extracting and processing the semi-structured data of web-systems has become relevant in the context of the information searching using Internet. In this paper the extracting of semi-structured data of web-systems has been described. The problems of the obtaining and classification of structured data from web-pages have been highlighted. Methods for solving the problems of the obtaining and classification of structured data from web-pages have been considered. The traditional models of knowledge representation are showed. The process of data extracting using the query language to the markup language elements is characterized. An algorithm for the web-sources, from which data will to be obtained, ontological model integrating creating is proposed.

References

1. Bondarenko M.F., Shabanov-Kushnarenko Yu.P.: Theory of intelligence. Textbook, X.: Izdvo SMIT, 576 p. (2007).
2. Buileaar P., Eigner T.: Topic extraction from scientific literature for competency management. In The 7th International Semantic Web Conference PICKME 2008, Karlsruhe, Germany, 55-67. (2008)
3. Kolada A.S., Gogunsky V.D.: Automation of information extraction from the science-computer databases, Management of the development of complex systems, No. 16 (2013)
4. Kushniretska I., Kushniretska O., Berko A.: Designing of Structural Ontological Data Systems Model for Mash-UP Integration Process, Applied Computer Science, 11(1) (2015)
5. Kushniretska I., Kushniretska O., Berko A.: The ontological model of knowledge of scientific and technical information system, Computer Science and Information Technologies (CSIT'2014): proc. of the IX-th Intern. Scientific and Techn. Conf., Lviv, Ukraine / Min. of Education and Science of Ukraine (2014)
6. Kushniretska I.: Semi-structured data dynamic integration Mashup system, Computer Science and Information Technologies (CSIT'2016): proc. of the XI-th Intern. Scientific and Techn. Conf., Lviv, Ukraine, Min. of Education and Science of Ukraine, 220-221 (2016)
7. Manning C., Raghavan P., Schütze H.: Introduction to Information Retrieval, Cambridge University Press, ISBN 0-521-86571-9, (2008).
8. Lytvyn, V., Pukach, P., Bobyk, I., Vysotska, V.: The method of formation of the status of personality understanding based on the content analysis. In: Eastern-European Journal of Enterprise Technologies, 5/2(83), 4-12 (2016)
9. Kravets, P.: The game method for orthonormal systems construction. In: The Experience of Designing and Application of CAD Systems in Microelectronics (2007).
10. Lytvyn, V., Vysotska, V, Veres, O., Rishnyak, I., Rishnyak, H.: Content linguistic analysis methods for textual documents classification. In: Computer Science and Information Technologies, Proc. of the XI-th Int. Conf. CSIT'2016, 190-192 (2016)

11. Zhao Li, Wee Keong Ng, Aixin Sun: Web data extraction based on structural similarity, Journal Knowledge and Information Systems archive, Vol. 8, Issue 4, 438-461 (2005)
12. Zhou L.: Ontology Learning: State of the Art, Information Technology and Management, 8 (3), 241-252 (2007)
13. Chen, J., Dosyn, D., Lytvyn, V., Sachenko, A.: Smart Data Integration by Goal Driven Ontology Learning. In: Advances in Big Data. Advances in Intelligent Systems and Computing. – Springer International Publishing AG 2017. P. 283-292 (2017).
14. Su, J., Vysotska, V., Sachenko, A., Lytvyn, V., Burov, Y.: Information resources processing using linguistic analysis of textual content. In: Intelligent Data Acquisition and Advanced Computing Systems Technology and Applications, Romania, 573-578, (2017)
15. Vysotska, V., Chyrun, L., Chyrun, L.: Information Technology of Processing Information Resources in Electronic Content Commerce Systems, CSIT, 212–222 (2016)
16. Vysotska, V., Hasko, R., Kuchkovskiy, V.: Process analysis in electronic content commerce system. In: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT 2015, 120-123 (2015)
17. Vysotska, V.: Linguistic Analysis of Textual Commercial Content for Information Resources Processing. In: Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET'2016, 709–713 (2016)
18. Basyuk, T.: The Popularization Problem of Websites and Analysis of Competitors. Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, vol 689. Springer, Cham pp. 54-65 (2017)
19. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Germany: LAP LAMBERT Academic Publishing (2016).
20. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
21. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
22. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
23. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
24. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
25. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
26. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
27. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
28. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, (2017)
29. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)

30. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
31. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
32. Hamon, T.: Biomedical text mining. In: Computational Linguistics and Intelligent Systems, colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf
33. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: Computational Linguistics and Intelligent Systems, COLINS, (2017)
34. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st International Conference Computational Linguistics and Intelligent Systems, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)
35. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)
36. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
37. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
38. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)
39. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
40. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf. (2017)
41. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
42. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)
43. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
44. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, (2017)
45. Kirkin, S., Melnyk, K.: Intelligent data processing in creating targeted advertising. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 131–132 (2017)
46. Hordienko, H., Ilchenko, M.: Development and computerization of an English term system in the fields of drilling and drilling rigs. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 129–130 (2017)

Influence of Readability on Popularization of Internet Resources

Taras Basyuk

Lviv Polytechnic National University, Lviv, Ukraine

Taras.M.Basyuk@lpnu.ua

Abstract. the article analyzes the influence of readability of the content on the popularization of Internet resource. The analysis of known methods for content readability determination, as additional component of popularization has been carried out. In researching, an algorithm for determining the basic level of readability has been developed, and recommendations for its application have been provided in order to enter the TOP results query. The question of determining the parameters of graphic design of the content is considered and recommendations on their application in the process of designing the Internet resource are provided. The research conducted creates the preconditions for further study of the subject and formation of mathematical support for automatic content rewriting depending on the required level of readability.

Keywords: readability; Internet resource; popularization; relevance; search engines.

1 Introduction

At the current stage of the development of digital technologies, there is a need to create effective websites, forums, and social networking pages that promote the product or company and attract new customers. Every day, more and more research market experts view the online resource as both mobile office and exhibition space, which requires their popularization in order to enhance the visibility of the company in the global network. As known [1,2], in the popularization process of the Internet resource the optimization of text content for search queries is one of the criteria for success. At the same time, "specialized texts" are created, which are optimized for the keywords and phrases-copywriting, which is divided into copywriting and SEO copywriting [3,4].

The objective of copywriting is to write genuine author's texts, and few special requirements are put forward to SEO-copywriting. Namely, texts should be not only original and unique, but also understandable to the target audience, properly structured and contain the necessary number of keywords. In addition, the number of keywords should clearly meet the requirements of search engines and readability

criteria, in order to avoid filtering [5]. SEO-copywriting plays an important role in the process of ranking an online resource. However, in practice, most of the texts placed on the pages are often written exclusively for search engines, which negatively affect company as well as goods and services that it provides. Considering that users observing high resource ratings are forced to look for less "popular", but more "valuable" sites for them.

Change of this situation is in the application of a variety of methods by SEO experts: increase of usability, reducing the time to load the resource, adding advertising, etc. However, the content and ways of submitting it was the main problem, which directly affects the ease of texts reading - readability. Readability is the feature of a text material that characterizes the ease of perception by a person [6,7]. Readability is a complex indicator, which includes: a clear match of keywords to the subject of the online resource and aesthetic design (from the background of the document to typography). Therefore, the actual task is to analyze the methods of assessing the readability of the content and determine their influence on the process of popularizing the Internet resource.

1.1 Analysis of recent researches and publications

The analysis of well-known statistical information services (Google Analytics, LiveInernet, HotLog, Open-Stat) showed that if the average time to browse the Internet resource page is less than ten seconds, then the latter does not meet the requirements of the user [8]. The main reason for this fact is the irrelevance of information about the user's request, that is, the web rob during the analysis made an error while analyzing the content of the Internet page. In order to identify possible causes, a well-known analysis of known types of relevance was carried out. Search engines are constantly implementing special algorithms to improve the quality of search results and display relevant information to the user. In this case, the extent to which search results match the task set in the search query is understood as the relevance of SEO, that is, the set of keywords or phrases contained in the text of the web site and their relevance, from the standpoint of the search engine, to user queries is determined [9-17]. The conducted analysis of literary sources showed that today the following types of relevance are distinguished as follows: formal, content-related and pertinent [18-31].

Formal - the main type, on which until recently mechanisms for ranking search engines were built. It algorithmically compares the search query model with the model of the document that was staged by the search engine. Namely, the "higher frequency" of the query text is found in the document of the given web site and rarer it occurs in other documents, the higher the weight of this resource on a separate request. It can be concluded that the words from the query should be inserted into the text of the page; the inadmissibility of excessive density of key phrases so that the resource does not enter into the search engine filters; minimizing rarely used phrases.

Content-related - relevance, which is determined by the non-formal approach and used by search engines for the assessment of the search quality. Namely, specialists of the search engine (assessors) form a conclusion regarding the compliance of user's

query information found by the search engine. Basically, this type of relevancy is used to enforce restrictions when accessing the plurality of "unwanted" resources.

Pertinent – is a degree of satisfaction with the results of search retrieval on the part of the user, that is, the degree of satisfaction with the information received by the expectations of the resource visitor. This type is the highest level of relevancy, more and more search engines are trying to provide it every day. In pertinent relevance, not only keywords or usability of the resource, but actual relevancy and readability of the information play an important role. The latter indicator is often left out of the attention of the developers of online re-source, although it creates the first impression and preconditions for the abandonment of the resource [32-45].

The conducted analysis shows that since the relevance of information is a subjective factor that dynamically varies depending on the subject area, special attention should be paid to the actual readability of the content. This parameter has a key value when determining the affinity relevancy. Therefore, the actual task is to analyze the impact of readability on the popularization of Internet resources in the Global Internet [11-26, 46-51].

1.2 Formulation of the problem

In general, the readability index does not fit into any analytic web metric of search engine. However, the latter is able to record both the failure rate and time spent on the page and depth of viewing (scrolling). The information obtained is used to assess the informativeness and readability of the content. In view of this, the purpose of the study is to analyze the methods that determine the readability of the content and determine their efficiency in the process of content evaluation. In order to achieve the goal set it is necessary: to carry out an analysis of known approaches to the definition of readability; develop a general algorithm for readability evaluating; formulate recommendations on the design of the text content of the Internet resource. The solution of the tasks set will provide the necessary tool for creating high-quality content and displaying it for the end user, and thus increasing the rating of the resource in results query.

2 Major research results

In the process of creating and publishing Internet resources, the most labor-intensive process is to write content that is mainly targeted to certain results queries. At the same time, most SEO-optimizers do not pay much attention to such a concept as "readability", since it is believed that if the page is in the top of the results query, then it is provided with "continuous flow of visitors". However, this situation may change after the next upgrade of search engine algorithms. Since the global transition to pertinent relevance is avoidless. Given that, taking into account the level of readability of the audience the number of visitors can be significantly increased and Internet resource in the global network can be popularized. The importance of determining the level of readability is confirmed by the study of excellent subjects of

Internet resources with the help of free readability.io service. This resource enables getting general information (readability index) of a separate network site. The following five most popular subjects in Ukraine were chosen to be studied: building materials, computer technology, toys, real estate, banking. The analysis consists in exploration of the most popular resources that form the TOP 5 in Google's search and averaging results to determine the level of readability. The results of the analysis are shown in Fig.1

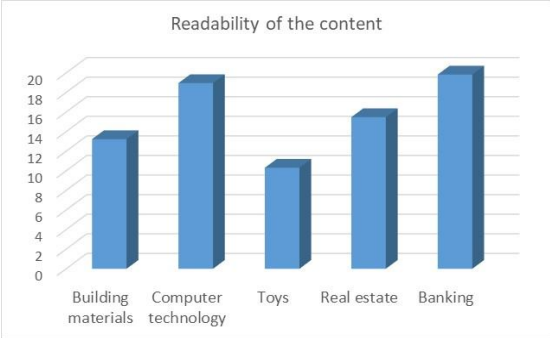


Fig.1. Readability of the content depending on the subject area

As can be seen from Fig.1, the readability of content is directly related to the topic, namely, the more complex it is, the higher the level of readability is needed, and therefore the resource that pretends to be in the TOP positions should correspond to the baseline. In order to conduct a comprehensive study of readability, five most popular methods were analyzed [6]: Flesch Reading Ease Scale; Canning Fog Index; Coleman-Liau Index; SMOG grade; Automated Readability Index.

Each of the methods pays considerable attention to the of parameters set in determining the readability of the content, so their detailed analysis was carried out in order to create a common approach that could be used in the process of popularizing Internet resources among relevant topics.

Flesch Reading Ease Scale is the scale of distribution of each particular index element, thus the higher the value the simpler text for perception is [6]. Namely, 80 – 100 - primitive level, 60-79 - simple texts, usually periodicals, 50-59 - fiction, 30-49 - level of business literature, professional and industry publications, up to 30 - level of scientific literature. The dependence based on which the index is determined for the Slavic languages has the following form:

$$FRE = 206.835 - (1.3 \cdot ASL) - (60.1 \cdot ASW), \quad (1)$$

where, *ASL* is average sentence length; *ASW* is average number of syllables per word. For correct work it is necessary to analyze the text fragment consisting of a minimum of 100 words.

Canning Fog Index – is used as an indicator in determining the level of readability required by the audience to understand the content. For Slavic languages it is characterized by the following dependence:

$$FI = 0.4(0.78 \cdot NWS) + (100 \cdot NWT), \quad (2)$$

where, *NWS* is the average number of words in a sentence; *NWT* is average number of words with length of syllables more than four; *0.78* is correction ratio for Cyrillic text.

Coleman-Liau Index is used in the process of mechanical assessment of the complexity of texts. Unlike the previous indicators, as parameters, not the number of syllables, but the actual letters, which can greatly simplify the process of counting them is used. Dependence has the following form:

$$CLI = 0.0588 \cdot L - 0.29 \cdot S - 15.8, \quad (3)$$

where, *L* is average number of letters per 100 words; *S* is average number of sentences per 100 words.

SMOG grade. The basic idea is that the complexity of the text increases with the raise in the number of complex words (words with three or more syllables) included in it. In general, it is an estimate of the dependence of complex words on the number of sentences and has the following form of dependence:

$$SMOG = 1.0430\sqrt{NP \cdot \frac{30}{NS}} + 3.1291, \quad (4)$$

where, *NP* is number of polysyllables; *NS* is number of sentences.

Automated Readability Index is measure of determining the complexity of the perception of text by the reader. According to this method, all texts are divided into 14 levels of readability (from the simplest to the more complex), which is determined according to the following dependence:

$$ARI = (4.71 \cdot CS) + (0.5 \cdot WS) - 21.43, \quad (5)$$

where, *CS* is the ratio of the number of letters to the number of words in the text; *WS* – the ratio of the number of words to the number of sentences in the text.

Conducted analysis of methods allowed developing a comprehensive algorithm for determining the level of readability with the use of statistics services. The block diagram of the algorithm is depicted in Fig. 2.

Step 1. Determination of the level of readability of subjects related to Internet resources. As it was shown, content readability should be commensurate with the audience of the users for which it is designed. In order to determine the required level of content, the comparison of the resources of the leading Ukrainian higher education institutions (in regions) according to the Google search engine rankings was made. The comparison was carried out by using the Readable.io service applied by developers to evaluate the readability of the content. The compliance with the readability level required for the preparation of respondents applied in Readable.io. is introduced in table 1.

Table 1. Meaning of readability levels

Readability level	Level of preparation	Readability	Level of respondent
A Level	Initial	Up to 10	High school (1-4 grades)
B Level	Below average	Up to 16	Average harm (5-10 grades)
C Level	Average	Up to 20	Average harm (11-12 grades)
D Level	Above average	Up to 24	High school
E Level	High	More than 24	Postgraduate study, PhD

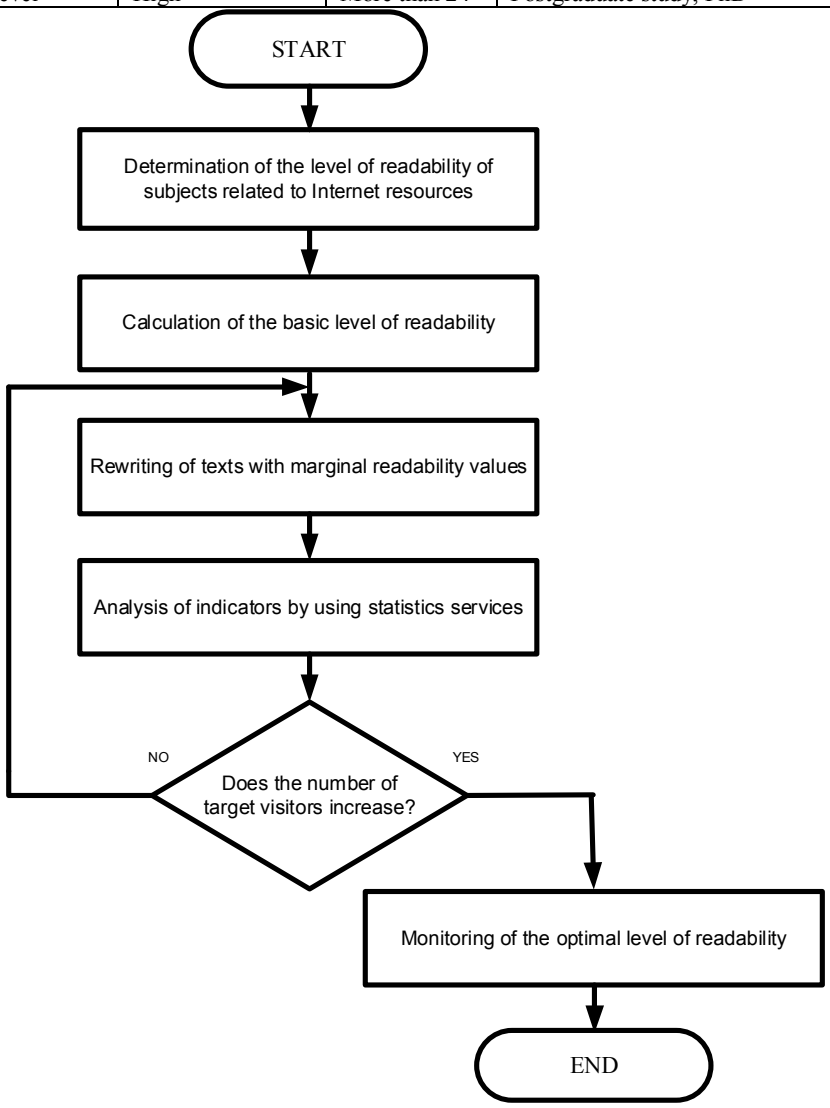


Fig.2. Algorithm for determining the level of readability

The analysis of resources of higher educational institutions showed the following results: Ivan Franko National University of Lviv (www.lnu.edu.ua), average readability - 16.3; Kyiv Taras Shevchenko National University of Kyiv (http://www.univ.kiev.ua), average readability - 25.7; National Technical University of Ukraine " Igor Sikorsky Kyiv Polytechnic Institute" (http://kpi.ua), average readability - 22.5; National Technical University "Kharkiv Polytechnic Institute" (http://www.kpi.kharkov.ua), average level of readability - 23,8; Kharkiv National University of Radio Electronics (http://nure.ua), the average level of readability - 18.3.

Step 2. Calculation of the basic level of readability (RBL). It is defined as the average result among other resources of the subject. Consequently, the baseline for this subject is determined as the arithmetic mean of the following indices, that is: *RBL=21,32.*

Step 3. Rewriting texts with marginal readability values. This stage lies in the processing of the text content of the Internet resource, with the maintenance of the content load, placed on the boundary of the base level. For comparison, the resource of the Lviv Polytechnic National University (http://lp.edu.ua) was analyzed, which received the following indicators: Flesch Reading Ease Scale (21,4); Canning Fog Index (24.6); Coleman-Liau Index (18.3); SMOG grade (20); Automated Readability Index (17.8). The average readability level is 20.4. As it is seen from the analysis, resource readability rating is defined as D, which is associated with the content of above the average complexity and necessary training of respondents at the level of higher school. The average level ranges within 5% of the basic level of the industry, which is the best indicator. Hence, the resource fully meets the criterion of readability in this topic. If the index has a deviation of more than 10%, then further manipulations are necessary, namely, "adjusting readability to the base level".

Step 4. Analysis of indicators by using statistics services. Even if the resource meets to the base level, it is necessary to increase the readability index and analyze the dynamics of time spent by users on the resource as well as interaction with it. If trends are positive, then there is an increase in readability, otherwise - on the contrary.

Step 5. Monitoring of the optimal level of readability. Based on the data of the previous step, the optimal level of readability for the visitors, which is supported by the writing of new texts, is optimally found on the experimental path.

2.1 Content graphic design guidelines

As the study showed, readability of content is important in the process of popularization. However, there is a situation when the text completion is executed according to all the rules, but when switch to the resource, the user still leaves it and the case relates to graphic design. The analysis of popular internet resources made it possible to determine the parameters and formulate recommendations that have to be followed in the process of popularization:

Font face and point size of the stick. It has the greatest impact on the readability of the text, and therefore requires the use of fonts without serifs, in particular Arial, Verdana, Tahoma and others. Because these types of fonts are correctly displayed on

all types of devices that run on popular operating systems. Optimum in terms of readability is the point size of 12pt. Fonts of greater or lesser size slow down reading and increase the fatigue of visitors.

Text justification and strings length. The conducted analysis showed that it is advisable to justify texts on the left edge, since right edge raggedness help users to focus their eyes by increasing both the speed of reading and improving perceptions of information. As to the length of the lines, it is experimentally proved that text columns having a width of more than 50-70 characters are difficult to read due to the variety of output means. As proof of this recommendation, one can consider the resources of the popular online editions: Ukrainska Pravda (https://www.pravda.com.ua), Lvivskaja Gazeta (http://gazeta. lviv.ua), Vysokiy Zamok (https://wz .lviv.ua) and others.

Use of uppercase. It is necessary to minimize the use of text paragraphs in uppercase. The studies show that this text is much more difficult to read on the one hand, and on the other hand, the upper case in Internet communities determines an increase in the tone of the presentation of the material.

Underlying in the text. One should not do underlying in regular text, since this indention is reserved for hyperlinks.

Contrast between text and background. There is a direct relationship between the contrast between text and background and readability. Namely, the high contrast of the latter increases the readability. In particular, for surfaces that reflect light (paper), the best combination is black text and white background. Although for surfaces emitting light the dependence is the opposite. However, due to polygraphy, dark text on the background became actually the standard for computer display. As an alternative, soft blend between text and background can be used, herewith the contrast between font and background should not to exceed 4:1.

The detailed analysis of both the methods for assessing readability and the content design parameters has shown that in order to enter the TOP of the search engine, in addition to the set of factors (relevancy, backlinks, domain, authority, etc.), it is necessary to follow the rules of texts writing. Namely, meet the expectations of users above all, concerning the requirements of readability and relevancy. This further will provide additional resource potential, both from users and search engines.

3 Conclusion

As a result of the research, direct dependence between the content readability and subject area was determined and the need for scientific research in this field was shown. The analysis of methods for determining readability has shown that each of them has its own peculiarities in application, and as to Slavic languages, the use of several methods with subsequent averaging of the results is the best tool. In view of this, an algorithm for determining the level of readability of the content of the Internet resource and method for its application was developed. The formation of recommendations for the graphic design of content, as an integral part of readability became the final stage of the study. The use of the described mechanisms will provide

additional means for popularizing the Internet resource and, accordingly, increasing the percentage of "satisfied" users.
Further research will be focused on the software development of intellectual content analysis for promoting online resources.

References

1. Grappone, J: Search Engine Optimization (SEO): An Hour a Day. United States: Wiley Publishing, Inc. (2013)
2. Basyuk, T.: The Popularization Problem of Websites and Analysis of Competitors. In: Advances in Intelligent Systems and Computing, 689. Springer, 54-65 (2017)
3. Thackston, K.: SEO Copywriting Flow: Creating a Steady Stream of Rankings & Conversions. In: CreateSpace Independent Publishing Platform (2012)
4. Vysotska, V., Chyrun, L., Chyrun, L.: Information Technology of Processing Information Resources in Electronic Content Commerce Systems, Computer Science and Information Technologies, 212–222 (2016)
5. Sullivan, D.: FAQ: All About The New Google "Hummingbird" Algorithm. Available online at: https://searchengineland.com/google-hummingbird-172816
6. Bailin, A., Grafstein, A.: Readability: Text and Context, Palgrave Macmillan (2016)
7. Chen, J., Dosyn, D., Lytvyn, V., Sachenko, A.: Smart Data Integration by Goal Driven Ontology Learning. In: Advances in Big Data. In: Advances in Intelligent Systems and Computing. – Springer International Publishing AG, 283-292 (2017)
8. Mokalis, A., Davis, J.: Google Analytics Demystified, CreateSpace Independent Publishing Platform (2018)
9. Khor, T.: Keyword Relevance in Search Engine Optimization, GRIN Publishing (2014)
10. Bentz, B.: Relevance Raises Response: How to Engage and Acquire with Mobile Marketing, SkillBites (2016)
11. Su, J., Vysotska, V., Sachenko, A., Lytvyn, V., Burov, Y.: Information resources processing using linguistic analysis of textual content. In: Intelligent Data Acquisition and Advanced Computing Systems Technology and Applications, Romania, 573-578 (2017)
12. Vysotska, V.: Linguistic Analysis of Textual Commercial Content for Information Resources Processing. In: Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET, 709–713 (2016)
13. Vysotska, V., Hasko, R., Kuchkovskiy, V.: Process analysis in electronic content commerce system. In: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 120-123 (2015)
14. Davydov, M., Lozynska, O.: Mathematical Method of Translation into Ukrainian Sign Language Based on Ontologies. In: Advances in Intelligent Systems and Computing, 689, 89-100 (2018)
15. Davydov, M., Lozynska, O.: Linguistic Models of Assistive Computer Technologies for Cognition and Communication. In: Computer Science and Information Technologies, Proc. of the Int. Conf. CSIT, 171-175 (2017)
16. Mykich, K., Burov, Y.: Uncertainty in situational awareness systems. In: Modern Problems of Radio Engineering, Telecommunications and Computer Science, Proc. of the Int. Conf. TCSET, 729-732 (2016)

17. Mykich, K., Burov, Y.: Algebraic Framework for Knowledge Processing in Systems. In: Advances in Intelligent Systems and Computing, Springer, 217-228. (2017)
18. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Germany: LAP LAMBERT Academic Publishing (2016).
19. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
20. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
21. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
22. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
23. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
24. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
25. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
26. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
27. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. . In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 100–109 (2017)
28. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)
29. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
30. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
31. Hamon, T.: Biomedical text mining. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf (2017)
32. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 47–55 (2017)
33. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)
34. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)

35. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
36. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
37. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)
38. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
39. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf. (2017)
40. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
41. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)
42. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
43. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 133–134 (2017)
44. Kirkin, S., Melnyk, K.: Intelligent data processing in creating targeted advertising. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 131–132 (2017)
45. Hordienko, H., Ilchenko, M.: Development and computerization of an English term system in the fields of drilling and drilling rigs. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 129–130 (2017)
46. Gorbachov, V., Cherednichenko, O.: Improving communication in enterprise solutions: challenges and opportunities. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 127–128 (2017)
47. Didusov, V., Kochueva, Z.: Statistical methods usage of descriptive statistics in corpus linguistic. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 125–126 (2017)
48. Verbinenko, Yu.: Discursive units in scientific texts. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 120–123 (2017)
49. Titova, V., Gnatchuk, I.: Evaluation ofa formalized model for classification of emergency situations. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 110–119 (2017)
50. Kuprianov, Ye.: Semantic state superpositions and their treatment in virtual lexicographic laboratory for spanish language dictionary. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 37–46 (2017)
51. Kotov, M.: NLP resources for a rare language morphological analyzer: danish case. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 31–36 (2017)

Using Dynamic Neural Networks for Server Load Prediction

Pavlo Pukach[1], Volodymyr Hladun[2]

Lviv Polytechnic National University, Department of Applied mathematics

[1]pavlopukach@gmail.com, [2]v_hladun@yahoo.com

Abstract. In this paper, the approach of using neural networks for making time series predictions of strongly nonlinear data is used. A brief examination of neural networks usage for time series predictions is given, as well as the definition and schematics of LSTM blocks. The comparison between the time delay values and the prediction accuracy is given. It is shown that certain values of time delay can greatly increase the prediction accuracy.

Keywords: neural networks, long short-term memory, time series prediction;

1 Introduction

Nowadays, there is a considerable amount of web applications running all over the world. In a classical web app architecture, a physical or virtual machine, called a node, is required for running the application. The node's characteristics, such as CPU and RAM amount directly affect the performance of the application – namely the amount of requests it is able to serve concurrently. Most applications require a single node for their deployment, but when the amount of its users gets significantly higher, a single node is not enough for optimal application performance.

Modern top-tier web apps rely greatly on scaling. They run on a cluster of nodes, with each node containing a single instance of the application, behind a load balancer tool, which decides what requests should be served by particular nodes depending on their actual workload. The issue with this setup is that the number of requests (the application or server load) varies over time. Thus, for effectively serving requests to end users, the system has to increase or decrease the number of nodes depending on user activity.

The goal of this paper is to use dynamic neural networks (DNN) for predicting server load. Accurate predictions can ensure that the node count is close to optimal, so there is no over-usage of computing time. This can greatly reduce application hosting costs, especially with cloud providers like Amazon EC2, which give a possibility to request nodes on-demand.

1.1 Neural Networks

Neural Networks approach is usually involved in time series predictions in which traditional prediction may not be able to capture the non-linear pattern in data[1].

Neural Networks can be classified into two categories: static and dynamic. Static (feedforward) networks have no feedback elements and contain no time delay. In another words, the output is calculated directly from the input through the connections. In dynamic network, the output depends not only on the current input, but also on the previous inputs to the network or estimated output of the network [2]. In this paper, we will use LSTM networks.

Long short-term memory (LSTM) units (or blocks) are a building unit for layers of a recurrent neural network (RNN). A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. Each of the three gates can be thought of as a "conventional" artificial neuron, as in a multi-layer (or feedforward) neural network: that is, they compute an activation (using an activation function) of a weighted sum. Intuitively, they can be thought as regulators of the flow of values that goes through the connections of the LSTM. There are connections between these gates and the cell.

The **time delay** (TD) is the number of previous network outputs that are taken into account for prediction of the next value.

2 Server load prediction

For the purposes of predicting server load based on time series, we conducted several experiments, using LSTM neural networks. The sample dataset that was used in this work, is derived from the request count logs on a AWS EC2 container. The numbers represent the web application request count per 5 minutes. The size of the dataset is 2016 entries, representing roughly one week of application request count logging.

Several LSTM neural networks were developed, with different values of time delay d – 1, 25 and 30.

We have trained the three networks on 90% of the sample dataset, and made test predictions on the rest of it. To compare the results, the RMSE (root mean square error) metric is used:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2}$$

3 Conclusion

Comparing the results, it is shown that the prediction accuracy varies depending on the time delay. For the first neural network, it is seen on the graph (Figure 1), that the network does not take into account the peaks of server load. The second and the third networks have greater values of time delay, and their predictions (Figure 2, Figure 3)

are way more accurate. It is also shown that all of the networks do not predict the two high peaks at and t=159. These peaks might represent a DOS-attack or some other abnormal cases, and such anomalous behavior is extremely hard to predict. However, the overall prediction results tend to get the more accurate the more the value of the time delay is. It follows from here, that dynamic NN's with certain values of time delay can make better predictions of strongly nonlinear data, such as server load.

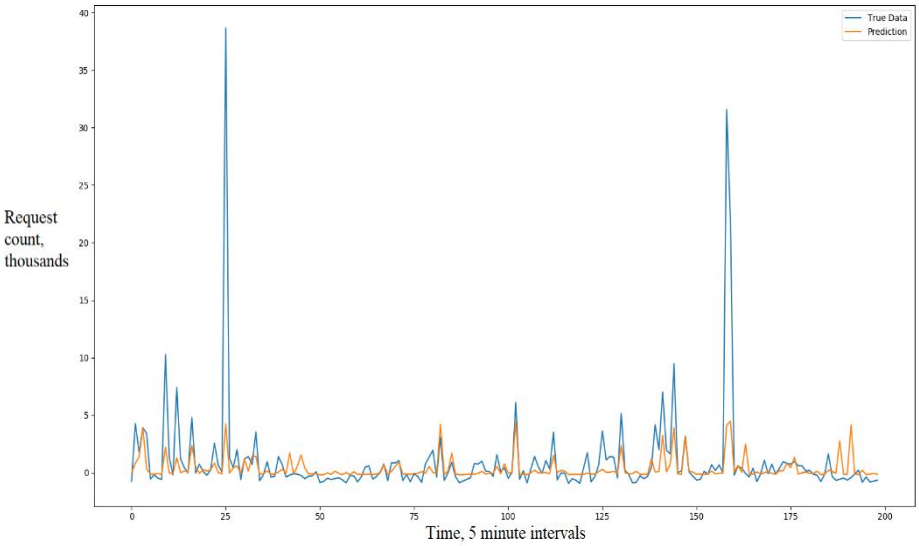


Fig. 1. Load prediction (d=1, $RMSE$=3.60)

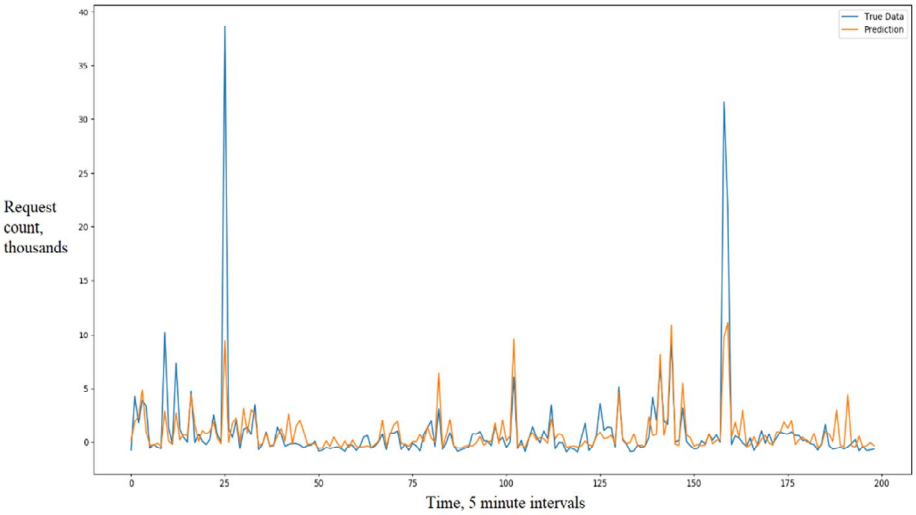


Fig. 2. Load prediction (d=25, $RMSE$=3.02)

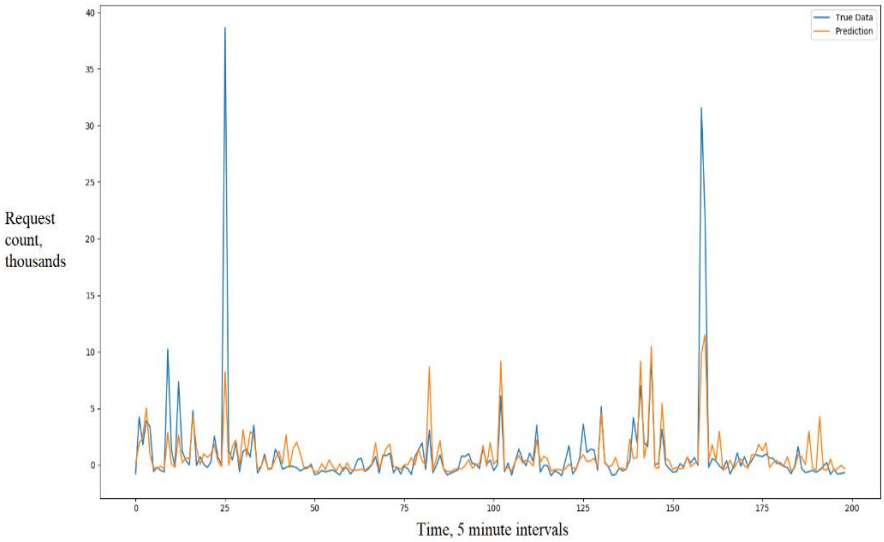


Fig. 3. Load prediction (*d*=30, *RMSE*=2.95)

References

1. M. Beale, M. Hagan, and H.Demuth, "Matlab neural network toolbox user's guide," The Math Works Inc., 2010, http://www.mathworks.com/help/pdfdoc/nnet/nnet ug.pdf.
2. C. A. Mitrea, C. K. M. Lee, and Z. Wu, "A comparison between neural networks and traditional forecasting methods: A case study," in International Journal of Engineering Business Management, vol. 1, no. 2 2009, pp. 19–24.

Peculiarities of Remote-Piloted Vehicles On-Board Navigation Complex Construction

Mykola Mykyichuk [1] and Volodymyr Markiv [2]

Lviv Polytechnic National University, Lviv, Ukraine

[1]mykolamm@ukr.net, [2]vovamarkiv6230@gmail.com

Abstract.. The article dwells upon the peculiarities of on-board navigation complex construction. It is highlighted that the optimal method for constructing on-board navigation complex is integration into single complex of sensors and systems with the integration of measurement information. The core of on-board navigation complex should be built on the basis of free-form inertial navigation system. To ensure the piloting tasks, the on-board equipment includes system of air signals. On the basis of the air signals system and magnetic compass air course counting is performed, which together with the inertial calculation allow to obtain comprehensive solution in an autonomous mode. It is important to include in the on-board navigation complex receiver of GNSS signals. Thus, the ideology of constructing the on-board navigation complex initially consists in the integration of measurements from the sensors and systems that make up its structure. It is emphasized that directly on-board navigation complex consists from inertial sensors, GNSS and magnetic compass receivers and also interface with air signal system. Specific types of sensors and systems are selected in accordance with the requirements of software and algorithmic support of on-board navigation complex.

Keywords: Remote-piloted vehicle, On-board navigation complex, Navigation system, System sensors.

1 Introduction

Robotic technologies are widely used in various types of air, land and sea transport, in agriculture, extraction of minerals and development of natural resources. At the same time, the market of robotic aircraft is developing very dynamically. As the total number of unmanned aerial vehicles grows, the task of integrating them into the common space with manned aircraft becomes urgent. It is possible only when the specified determination quality of remote-piloted vehicles parameters including precision and interference immunity is achieved. Operational standards are under development by the authorized civil aviation authorities and are likely to repeat similar requirements for on-board equipment of civil aircraft.

It is emphasized that total number of remote-piloted vehicles have increased so that the task of their integration into the common space with manned aircraft becomes urgent. It is possible only when the specified quality of the remote piloted vehicle movement parameters have been determined, including accuracy and interference immunity. It has been highlighted that remote-piloted vehicle equipment is subject to stringent requirements for minimization of cost, mass and size characteristics and power consumption, which are often mutually contradictory, and their implementation in general leads to deterioration of accuracy and interference immunity. The problem of ensuring accuracy and noise immunity when using the general-purpose element base becomes urgent.

Nowadays, international market do not offer specialized serial navigation systems for unmanned aircraft, especially small and medium-sized classes that meet the safety requirements in the general airspace. There are strict requirements for the on-board equipment of remote-piloted vehicle concerning minimizing of costs, weight and size characteristics and power consumption, which are often mutually contradictory, and their implementation in general leads to deterioration of accuracy and interference immunity. That is why the problem of ensuring accuracy and noise immunity when using the general-purpose element base is urgent [6,15,16,22].

To achieve this goal it is important to develop the concept of constructing the on-board navigation complex and increase noise immunity with variable structure.

2 Peculiarities of on-board navigation complex construction.

Specificity of remote-piloted application without terrestrial support under low-altitude flight, while reducing the visibility of signals of navigation satellites of global navigation satellite systems complicates the problem of accuracy and interference immunity provision.

The solution of the problem is possible in two ways [1,5]:

- use of the equipment used in manned aircraft (advantage of this approach is the use of waste products and technologies, disadvantage is ignoring the remote-piloted vehicle specificity that makes practically impossible to use it as a part of the remote-piloted vehicle of small and middle classes
- creation of specialized navigation systems of small and middle-class remote-piloted vehicle in which it is possible to apply inexpensive general-purpose sensors (possibility to maintain algorithms software, and apparatus core of remote-piloted on-board navigation complex based on to the inertial navigation system)

It is necessary to substantiate the concept of construction, development of software and algorithmic support and hardware solutions to improve accuracy, as well as to study the properties of on-board navigation complex with high interference immunity with variable structure for remote-piloted vehicles.

At the first stage of development of on-board navigation complex it is necessary to justify its structure in accordance with the requirements for the accuracy of determining orientation and navigation parameters, to propose the concept of the

development procedure, operation modes, to develop the structure of algorithms for complex information processing, orientation and navigation algorithms, justify the choice of the hardware of on-board navigation complex.

The basic requirements for the accuracy and range of measurements of orientation and navigation parameters of the on-board navigation complex correspond to the requirements of the inertial-satellite navigation system [5, 27].

The traditional approach to the design of the remote-piloted vehicles involves the following:

- selection of measurers
- development of orientation and navigation algorithms based on measurer values
- development of algorithms for complex processing of information.

The use of such an order is due to the small volume of sensors, high cost of products. However, more flexible and productive is the reverse order:

- requirements
- modes of operation
- measurements
- development of algorithms for complex processing of information.
- development of orientation and navigation algorithms based on measurer values
- selection of measurers

It is more convenient due to the lack of mandatory requirements for remote-piloted vehicle navigation equipment and, accordingly, the absence of the need for certification, which gives advantages in the development of algorithmic support and choice of measurers.

When creating remote-piloted vehicle, it is necessary to take into account such critical factors as ensuring the accuracy levels established by the technical specification while minimizing the cost, weight and size characteristics and energy consumption. From this point of view, the optimal method for constructing on-board navigation complex is integration into single complex of sensors and systems with the integration of measurement information. The core of on-board navigation complex should be built on the basis of free-form inertial navigation system. To ensure the piloting tasks, the on-board equipment includes system of air signals.

On the basis of the air signals system and magnetic compass air course counting is performed, which together with the inertial calculation allow to obtain comprehensive solution in an autonomous mode (without the use of GNSS). It is important to include in the on-board navigation complex receiver of GNSS signals. Thus, the ideology of constructing the on-board navigation complex initially consists in the integration of measurements from the sensors and systems that make up its structure. Directly on-board navigation complex consists from inertial sensors, GNSS and magnetic compass receivers and also interface with air signal system[2,3,4,15].

Specific types of sensors and systems are selected in accordance with the requirements of software and algorithmic support of on-board navigation complex.

It is based on the sequence of stages (requirements - operating modes - measurements - development of algorithms for complex processing of information - development of orientation and navigation algorithms based on measurer values - selection of measurers).

The navigation system of remote-piloted vehicle is depicted in Fig.1[3, 13, 18, 19].

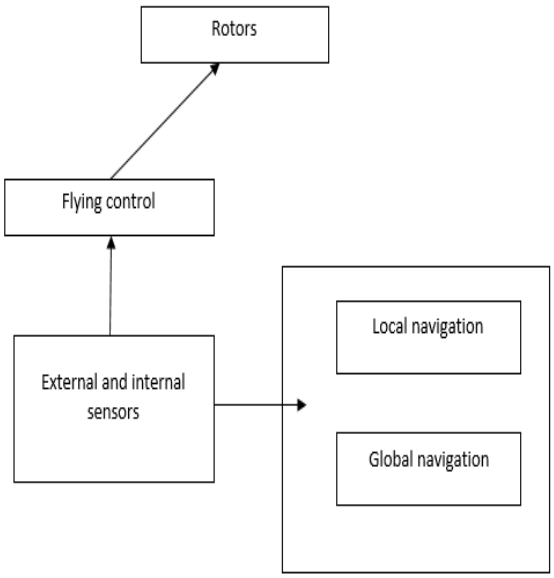


Fig. 1 Navigation system of remote piloted vehicle

Self-positioning and modeling of environment relative to the local coordinate system is under control of local navigation. 3D information is formed based on a set of flat images, forming 3D map of the studied surface. Block "Global navigation" presents the positioning of the remote-piloted vehicle by using the constructed 3D model of the external environment.

This technique allows to solve the problem of positioning in the condition of absence of GNSS signals. Local positioning system takes into consideration the onboard sensors indices, external sensors and position change data from the camera, which after processing by the filter of Kalman-Bucy solve the problem of orientation in space of dense urban development[7,8,9].

For the practical implementation of the proposed algorithm of the selection of a preferred profile of detection is necessary to ensure the adequacy of the technological process of production of photographs using the remote-piloted vehicle. For this purpose it was necessary to develop the methodology. Significant differences of this method consist in the use of on-Board 4-processor vector computer capable of processing images. The compute engine also provides automatic flight of the remote-piloted vehicle in the absence of GNSS data reception, focusing on the testimony of the inertial block, which includes a combination of accelerometers and gyroscopes.

For processing data from the sensors responsible for the positioning, was used an iterative formula to calculate the Kalman-Bucy coefficient[17, 20, 21, 24]:

$$E(e_{n+1}^2) = \frac{\sigma \frac{2}{k}(Ee_n^2 + \sigma_m^2)}{Ee_n^2 + \sigma_m^2 + \sigma_k^2}$$

where e_n^2 and e_{n+1}^2 – square of failures in n i n+1 moment of time accordingly; Ee_n^2 and $E(e_{n+1}^2)$ – expected value square of failures in n i n+1 moment of time accordingly; $\sigma_m^2 + \sigma_k^2$ – dispersion of positioning inertia block values and receiver of GNSS signals respectively.

By storing in the automated system of retrospective spatial data it is possible to set the flight mission, specifying only the name of the study area. By applying to the obtained areas data filter of Kalman-Bucy, it is possible significantly to improve the accuracy of the inertial navigation system due to the complete independence of measurement errors by these methods[23,25].

The use of the two-level scheme for the implementation of the prototype has allowed to implement a secure debug mode, because MultiWii allows the remote-piloted vehicle to hover or return to the starting point by the elementary route in the case of a failure in the block of "Remote-piloted vehicle navigation". The application in this mode of ultrasonic sensor allows to avoid collision with an obstacle by the elementary route. Operational adjustment of parameters takes place via radio or GPS in flight. At the time of the remote-piloted vehicle positioning on the landing pad are loaded the initial flight assignments and uploaded the collected information on the ground station via WiFi network.

The high speed of errors accumulation in inertial systems of positioning is caused by error of measurements of the used integrated accelerometers and the need to calculate the integral, leads to the multiplication of errors and does not allow to obtain positioning accuracy comparable with the accuracy of GNSS [8, 18,19].

The positioning of remote-piloted vehicle is made on the basis of the analysis of the location of ground objects and pre-designed models of environment. The block "Flying task and control" is responsible for the collection and processing of the 2D images obtained through the onboard camera, and solves such important tasks:

- adjustment of coordinates of an inertial positioning system;
- stabilization with respect to a predetermined position;
- safe landing of remote-piloted vehicle.

The on-board navigation complex has three main modes: "Initial parameters", "Navigation", " Attitude-and-heading reference" and four auxiliary modes: “Test Control", "Deviation", "Calibration", "Axis Coordination" (Fig. 2).

Modes "Deviation", "Calibration" and "Coordination of axes" can not work simultaneously and outside the mode " Attitude-and-heading reference ". In the "Test-control" mode, the other modes are not working.

In the "Navigation" mode on-board navigation complex generates and outputs navigation parameters with the required level of accuracy and in full volume. For successful operation of the regime, stable reception of GNSS signals and / or air

signal system is required. In the absence of reception of GNSS and / or air signal system during a time interval of a predetermined duration, on-board navigation complex switches to the operation mode of Attitude-and-heading reference. In this mode, according to the navigation parameters, there are signs of failure, and the UAV orientation parameters are determined with increased errors [26, 27,28].

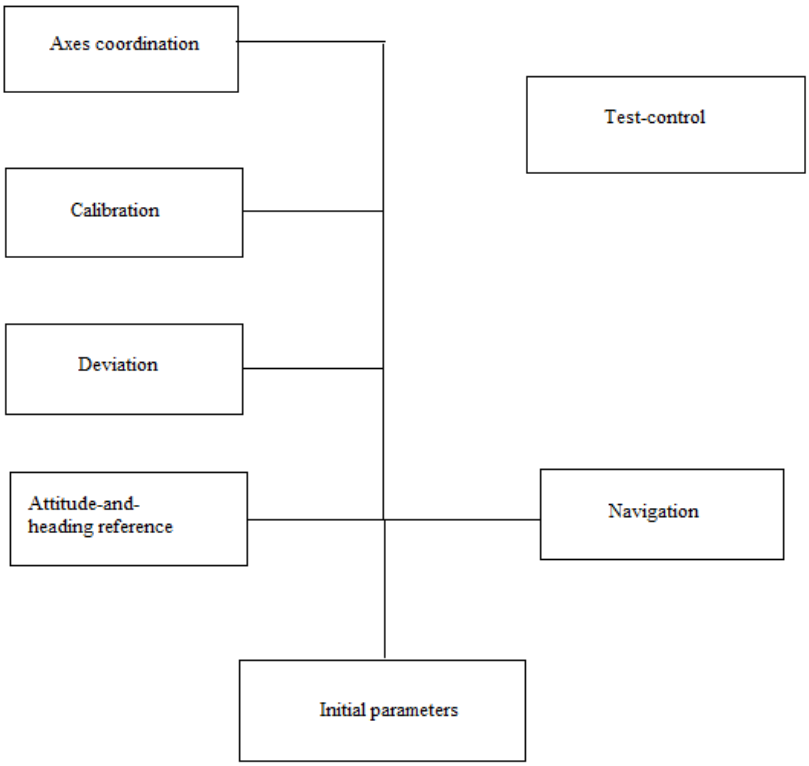


Fig.2 On-board navigation complex operation modes structure

In the Attitude-and-heading reference mode, the angular position of the remote-piloted vehicle relative to the vertical is determined by the measurements of the accelerometers. The error in determining the angles in this mode depends on the flight mode of the remote-piloted vehicle, the maximum accuracy is achieved with a flight close to a straight uniform. In the modes of intensive maneuvering of the remote-piloted vehicle, an error is accumulated, the magnitude of which depends on the duration and intensity of maneuvering. After the reduction of the effect of accelerations and angular velocities, the errors decrease. When start receiving GNSS signals again, the "Navigation" mode is restored. The proposed structure and logic of

switching modes fully meets the requirements of the on-board navigation complex and makes it easier to conduct routine maintenance[3,5].

3.1. On-board navigation complex hardware

It is necessary to determine the structure of on-board navigation complex. It includes sensors and systems that allow constructing complex that meets the minimum requirements for determining the orientation and navigation parameters: the inertial module, the satellite navigation receiver GPS and magnetic compass[10,11].

The output information of the inertial module are three projections of the absolute angular velocity of rotation and three projections of the apparent acceleration onto the orthogonal axes of the coordinate system. At the same time, it is required to provide a range of measured angular velocities.

The choice of the on-board navigation complex calculator is carried out taking into account the requirements for the interaction interfaces and the functional load [5, 12,13,14]:

- control, monitoring and signaling functions for all elements (basic and
- optional) that are part of the on-board navigation complex;
- collection, processing and transmission of data on internal and external data buses;
- functions for synchronizing the operation of sensors, systems and algorithms;
- implementation of algorithms for orientation, navigation and auxiliary algorithms;
- recording and storage in the non-volatile memory of initial and current settings of parameters of system.

Interfaces of on-board navigation complex interaction should provide reception and transfer of single discrete commands and data packets exchange with the onboard equipment. In on-board navigation complex should be implemented:

- four digital inputs: "Chassis compression" signal, "Axes coordination" command, "Test control" command, "Calibration" command;
- four discrete outputs: signals "Serviceability", "Readiness", "Power", "Error";
- channel for monitoring and controlling GNSS receiver parameters;
- channel for the exchange of data packets with on-board equipment;
- channel designed for high-speed transmission of measurements of inertial sensors, measurements of GNSS receivers in consumer equipment;
- channels designed for receiving and transmitting data to onboard equipment of the remote-piloted vehicle.

The "Chassis compression" signal is used to indicate the mode of stop or the movement of remote-piloted vehicle over the runway. "Axes coordination" is designed to provide a command for coordinating the axes of the magnetic compass. The "Calibration" command is used when calibrating inertial sensors. Integral signal "Serviceability" is formed on the basis of information about internal control of the status of subsystems, if one of the elements fails, the active signal level is not formed.

The "Readiness" signal is integral, the active state is formed when there is minimum information for measuring the implementation of one of the main modes and the end of the initial exhibition. The "Power" signal is used to indicate the presence of power. The "Error" signal is used when diagnosing the program error of the on-board navigation complex with built-in monitoring functions.

The proposed hardware structure allows to obtain the entire required volume of measurement information for the implementation of orientation and navigation algorithms.

4. Conclusions

Nowadays remote-piloted vehicles are widely used in different spheres of life, particularly in the area research. They have a lot of advantages: high economic efficiency; low altitude of aerial photoshooting; aerial photoshooting exactness.

But also here are some hindrances at any height, in particular, if take into account low-level remote-piloted vehicles. Then the flight is complicated due to the high turbulence of the atmosphere and the inability to track the change in the altitude, as on small unmanned aerials, the device for measuring the relative altitude is used for landing or absent altogether. Therefore, it is necessary to increase the height of the automatic flight and thus reduce the efficiency of the accomplishment of the tasks .

The conception of constructing on-board navigation complex is the following: determination of requirements, operating modes, selection of measurement types, development of algorithms for complex processing of information, development of orientation and navigation algorithms based on readings of measurers, the choice of measurers. This approach allows to get the desired result with the reduction of time and material costs.

The structure of algorithmic support of complex processing of information, solving problems of navigation and orientation of the on-board navigation complex, allowing to obtain the required characteristics while minimizing the cost of development and the product as a whole, is developed.

The hardware composition and interfaces of the interaction of the on-board navigation complex, which allow implementing the proposed algorithmic support structure, are proposed.

References

1. Борискин А.Д., Вейцель А.В., Вейцель В.А., Жодзишский М.И., Милютин Д.С.
2. Аппаратура высокоточного позиционирования по сигналам глобальных навигационных спутниковых систем: приёмники-потребители навигационной информации, 2010
3. Микийчук М., Марків В. Особливості системи управління безпілотними літальними апаратами, Матеріали V Міжнародної науково-практичної конференції “Математика. Інформаційні технології. Освіта, 69–71, 2017

4. Микийчук М., Марків В. Особливості GPS-спуфінгу щодо управління БПЛА, матеріали 6-ої Міжнародної наукової конференції ІКС-2017, 61-62, 2017
5. Харин Е.Г., Копелович В.А., Копылов И.А., Требухов А.В., Ларионов С.В. Результаты лётных испытаний интегрированной инерциально-спутниковой навигационной системы, 2014
6. Austin R. Unmanned aircraft systems UAVs design, development and deployment. - West
7. Sussex, PO19 8SQ, United Kingdom: John Wiley & Sons Ltd, 2010
8. Barton J.: Fundamentals of Small Unmanned Aircraft Flight. Johns Hopkins APL Technical Digest. V. 31, No. 2 , 132-149, 2012
9. Bond L.: Overview of GPS Interference Issues. GPS Interference Symp., Volpe National Transportation System Center, 28-32, 1998
10. Brown A.K., Yan Lu Performance Test Results of an Integrated GPS/MEMS Inertial Navigation Package / ION GNSS 17th International Technical Meeting of the Satellite Division, Long Beach, CA, 2004
11. Forssel, B. Olsen T.: Jamming Susceptibility of Some Civil GPS Receivers. GPS World, No. 1, 54-58. 2003
12. Grewal M.S., Weill L.R., Andrews A.P. Global Positioning Systems, Inertial Navigation, and Integration. – New York: John Wiley & Sons, Inc, 2001.
13. Kim J.-H., Sukkarieh S. Flight Test Results of GPS/INS Navigation Loop for an Autonomous Unmanned Aerial Vehicle (UAV) / ION GPS, 24-27 September 2002, Portland, OR, 2002
14. Key E.: Technique to Counter GPS Spoofing. Int. Memorandum, MITRE Corporation, 1995.
15. Lawrence A. Modern Inertial Technology (Navigation, Guidance, and Control). – New York:Springer-Verlag Inc, 1998.
16. Martin, M.: Non-linear DSGE Models and The Optimized Central Difference Particle Filter, 2-45, 2010
17. Markiv V.: Analysis of remote-piloted vehicles use and control system description". , Computer sciences and information technologies, No. 843, 347-351, 2016
18. Markiv V.: Justification of remote-piloted vehicles use and metrology supply improvement. 5th Int. Scientific Conf. ICS-2016, 20–21, 2016
19. Mykyichuk M., Markiv V. Metrology tasks of airphotoshooting by remote-piloted vehicle**,** Вісник "Радіоелектроніка та телекомунікації", В-во НУЛП.,№ 874**,** 57-61, 2017
20. Mykyichuk M., Markiv V. Peculiarities of fractal analysis of remote-piloted vehicles recognition, VI-a Міжнародна науково-практична конференція "Практичне застосування нелінійних динамічних систем в інфокомунікаціях.,. 20–21, 2017
21. Mykyichuk M., Markiv V. Peculiarities of the radio signals and hindrances in the navigation system of the remote-piloted vehicles, Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska, IAPGOŚ, № 8 (1), 40- 43, 2018
22. Neitzel, F., Klonowski, J.: Mobile 3d mapping with a low-cost UAV system. Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVIII-1/C22, 67-70.
23. Roach. D.: Dimensionality analysis of patterns: fractal measurements, Computers Geosciences, 1993, 849-869.
24. Salychev O.S. Applied Inertial Navigation: Problems and Solutions, BMSTU, 2004.
25. Sandau K.: Measuring fractal dimension and complexity - an alternative approach with an application, 164-176, 1993

26. Savage P. G. Strapdown Analytics Part1&2. – Maple Plain, Minnesota: Strapdown Associates,Inc, 2000.
27. Strang G., Borre K. Linear Algebra, Geodesy, and GPS. – USA, Wellesley: Wellesley-
28. Cambridge Press, 1997.
29. Tsui J. B.-Y. Fundamentals of Global Positioning System Receivers. A Software Approach. –Hoboken, New Jersey: John Wiley & Sons, Inc, 2005.
30. Vincenty T. Direct and Inverse Solution of Geodesics on the Ellipsoid with Application of
31. Nested Equations [Электронный ресурс] / Survey review.- Kingston Road, Tolworth, Surey,1975. – Режим доступа: http://www.ngs.noaa.gov/PUBS_LIB/inverse.pdf – 21.01.2015.
32. Winkler S., Schulz H.-W., Buschmann M., Vorsmann P. Testing GPS/INS Integration for
33. Autonomous Mini and Micro Aerial Vehicles / ION GNSS 18th International Technical Meeting of the Satellite Division, 13-16 September 2005, Long Beach, CA.

Study Methods of Image Segmentation for Intelligent Surveillance Systems

V.V. Hrytsyk

Lviv Polytechnic National University, Lviv, Ukraine

Abstract. The research of the machine-machine interface element is presented for the possibility of dynamic adaptation to improve the perception of the very least environment by developing a technique for adapting the robovision. The study include comparing between different threshold methods of segmentation of images of different categories.

Keywords: image segmentation, thresholds, future robots.

1 Introduction

Computer technologies are so important in the life of a modern society, that is hard to imagine any kind of activity where in one or another way does not use computers.

Persistent tendencies towards attempts to replace people with computers in all business processes where possible are in the world today [3]. The 4th industrial revolution taking place in developed countries stimulates the progress of knowledge, accordingly, the development and implementation of technology is sharply spurred every day, because the creation of some of the latest technologies serves as the foundation for the emergence of new ones. In many industries, science, Internet technologies, etc., developments related to the use of systems of artificial intelligence, as well as various types of robots or robotic systems are effectively used already. Overview of Research and Implementation of Intelligent IT in the EU and the United States: the EU finances the development of sophisticated techniques for understanding audiovisual life-style representations for typical and non-specific groups through the FP7 program [3] and continues to actively fund developments in this area (Horizon 2020). Therefore, the task of research of models and methods of synthesis of methods of perception taking of data of the visual spectrum obtained in real time working out for the machine-machine interface is actual. [2,4].

2 Problem

The perception of the conditions and patterns of the visual spectrum it is technologies with intellectual possibilities: they must feel and understand the real world dynamic [7]. Therefore, the development of computer vision systems with the dynamic perception of the external environment is a critical problem for the next generation tasks. Thus, author looking for best approach to build principles of computer

perception of the outside world through the understanding of video data [1-5,11]. The part of this problem it is the problem of image segmentation is considered in work.

3 Purpose

The aim of the work is to improve the perception of the visual input data by developing methods for adapting it to the features of the environment.

4 Experiments

In order to determine the most effective methods of segmentation, we study and take into account some of the most common methods of segmentation of the visual field of real-time knowledge. Author studied different mathematical approaches for processing of information. The approaches are taken from the following categories: (1) Histogram based Entropy computation methods; (2) Iteration based methods; (3) Histogram based modeling methods; (4) Histogram profile based methods; (5) Fuzzy methods.

4.1 Niblak method

The method is based on the brightness of local threshold calculation.

The method idea is to align the threshold of brightness binarization from point to point based on the deviation of the local average brightness value (the value calculated for each pixel based on the brightness values of itself and its neighbors), from the local (calculated for only one pixel) in the given mask [6]:

$$a_{xy}^{new} = \begin{cases} 0, \; if \; B(x,y) \le L \\ 1, \; if \; B(x,y) > L \end{cases}$$

where $B(x,y) \in [0,255] = \frac{r+g+b}{3}(a_{xy})$ is local brightness value of a_{xy} pixel; $L \in [0,255] = m_{w\times w}(x,y) + k \cdot s_{w\times w}(x,y)$ is local brightness value for pixel a_{xy} in the $w \times w$ neighborhood; $m_{w\times w}(x,y) \in [0,255] = \frac{\sum_1^{w\times w} B(x,y)}{w \times w}$ is average brightness value in the neighborhood of pixel $w \times w$; $s_{w\times w}(x,y) = \sqrt{\frac{1}{w \times w}\sum_1^{w\times w}(B(x,y) - m_{w\times w}(x,y))^2}$ is mean square deviation of the sample in the area (neighborhood) of the pixel; $k(const)$=-0,2 for objects if $B(x,y)\le 127$), and k=0,2 for objects if $B(x,y)>127$); $w(const)$ is neighborhood size, for example 16[6].

4.2 Bernsen method

The method is to find the threshold based on the values of the local maximum and the minimum brightness of the pixel

$$B(Avg_l) = \frac{Min(l) + Max(l)}{2} \text{ and } a_{ij} = \begin{cases} 0, if\ Avg \le const(l) \\ 1, if\ Avg > const(l) \end{cases}$$

where $Avg \in [0,255]$ is average value of brightness; Min_l *and* Max_l is minimum and maximum brightness value in the mask, respectively; a_{ij} ia new pixel brightness value; $l \in [0,255]$ is local level of brightness, the minimum and maximum value vary depending on the brightness that falls into the area.

4.3 Yen method

This method refers to methods that use the entropy of the distribution of the brightness of colors in the image. Yen's method looks at the object on the images and the background on which this object is located, as two different sources of visual information. The value of brightness, in which the sum of these two entropies reaches its maximum, is considered as optimum threshold for image segmentation [8].

At first step histogram of $p(l)$ image and incidence rate $N(l)$ of every level of image brightness is calculated. Next, we need sum brightness N_T of image pixels: $N_T = \sum_{i=0}^{\max(G)} p(i)$. Build helping histogram: $p_{norm}(i) = \frac{p(i)}{N_T}$,

$$p_{normC}(i) = p_{normC}(i-1) + p_{norm}(i),\ p'_{norm}(i) = p'_{norm}(i-1) + p_{norm}(i+1)^2,$$
$$p''_{norm}(i) = p''_{norm}(i+1) + p_{norm}(i+1)^2.$$

Find entropy of objects and background:

$C_{ob}(T) = -\log\{p_{normC}(i) \times (1 - p_{normC}(i))\}$, $C_b(T) = -\log\{p'_{norm}(i) \times p''_{norm}(i)\}$,

value i is calculated: $L = ar\,ma\,x_i\{C_b(T) + C_f(T)\}$.

L is used as threshold: $a_{xy}^{new} = \begin{cases} 0, if\ B(x,y) \le L \\ 1, if\ B(x,y) > L \end{cases}$

4.4 Minimal value method

The method of setting the threshold based on the minimum between two maxima. Maxims are the maximum values of the distribution of the brightness of the background image and the object itself on the histogram t=min($h[b_{ij}]$), where t is threshold value; b_{ij} is pixel brightness; $h[b_{ij}]$ is brightness distribution value on the histogram; min($h[b_{ij}]$) is minimal brightness distribution value on the histogram between two maxima.

4.5 Average value method

The method find average value based on image histogram: $t = \frac{\sum_{i=1}^{n} h[b_{ij}]}{n}$, where t is threshold value; n is quantity of histogram elements; $h[b_{ij}]$ is brightness distribution value on the histogram.

4.6 Triangle method

According to the method [9]: threshold – is element with maximal distance to line s from minimal b_{min} to maximal brightness value b_{max} on the histogram: t=max(d), where t is threshold value; d is distance from brightness value on the histogram $h[b_{ij}]$ to line value on the histogram (fig. 1).

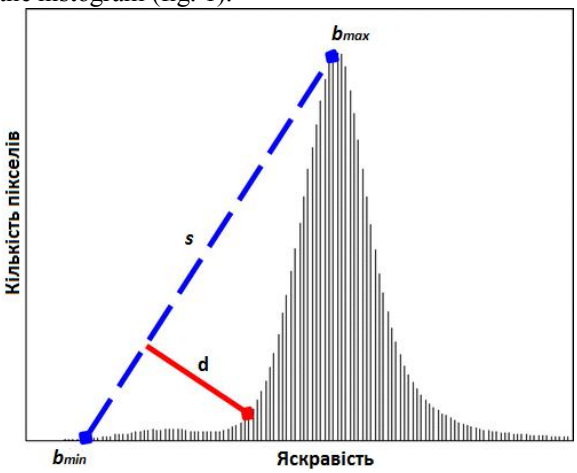


Fig. 1. Triangle method

$$L = argmax_{b(i)} d \ ,$$

where L is threshold value; d is distance from pixel brightness value $b(i)$ to the s.
Next, binary procedure by standard formula is realized:

$$a_{xy}^{new} = \begin{cases} 0, if \ B(x,y) \le L \\ 1, if \ B(x,y) > L \end{cases}$$

4.7 Otsu method

Method use histogram distribution of pixel brightness value of image. Histogram with value $p_i = \frac{n_i}{N}$ is built: $\omega_0(k) = \sum_{i=0}^{k} p_i$, $\omega_1(k) = \sum_{i=k+1}^{L} p_i = 1 - \omega_0(k)$, $\mu_0(k) = \sum_{i=0}^{k} \frac{ip_i}{\omega_0}$,

$\mu_1(k) = \sum_{i=k+1}^{L} \frac{ip_i}{\omega_1}$, $\eta(k) = max_{k=1}^{L-1}\left(\frac{\sigma_{cl}^2(k)}{\sigma_{all}^2}\right)$, where N is general image pixel quantity; n_i is i-level brightness image pixel quantity; $k \in Z, \in \overline{0, L}$; $\omega_0\omega_1$ is relative frequencies of the corresponding classes; $\mu_0\mu_1$ is average levels for each of the image classes; $\eta(k)$ is maximum value of image section quality; $\sigma_{cl}^2 = \omega_0\omega_1(\eta_1 - \eta_0)^2$ is inter-class dispersion; σ_{all}^2 is general dispersion of whole image.

5 Correlation is realized by MSE

$$MSE = \frac{1}{m*n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left|I(i,j) - K(i,j)\right|^2 ,$$

were $I(i,j)$ is pixel brightness value on the point (i,j) ; K and I have $m*n$ resolution.

6 Experiment results

Correlation MSE method was used. The method with the lower threshold of binary process was used as the reference image. Studied image was an image obtained as a result of the processing of the original one by each of the methods. The error was calculated by the next formula.

$$MSE = \frac{1}{w*h}\sum_{i=0}^{w-1}h\sum_{j=0}^{n-1}\left|I(i,j) - K(i,j)\right|^2 ,$$

where $I(i,j)$ is the value of the pixel's brightness at the point (i,j) of the under study image; where $K(i,j)$ is the value of the pixel's brightness at the point (i,j) of the standard image; w is image width characteristic; h is image height characteristic; MSE is mean squared error value.

Experiment results are in the table 1.

Table 1. Comparison of results

	1	2	3	4	5	6	7	8	9	Похибка
Global thresholds	9275	1032	2635	1916	1865	2128	1089	2799	2921	±248
Niblack's method	13581	16124	16388	16856	14841	11073	15014	3724	11600	±412
Bernsen's method	2494	2357	4869	2790	4891	3002	3406	3714	3887	±211
Yen	7049	9618	10518	7915	9765	10141	7360	11302	10295	±352
Method of minimal	7946	4351	2562	2791	1944	2248	4382	7216	3835	±339
Method of average	4261	5563	6783	4598	5999	6360	4272	7322	6450	±216
Triangle's method	7005	8723	9541	7625	9726	9952	7281	10713	9461	±285
Otsu	7670	984	2421	1871	1831	1952	1052	2519	2675	±231

The errors regarding the same pattern obtained from the evaluation of various methods in the table is reflected.

7 Conclusion

Using table 1 results, we can build adaptive to illumination system. System can use different threshold to different situation.

References

1. Грицик В. В. Оцінка якості передавання і комп'ютерна обробка даних образів / Грицик В. В. // Доповіді НАН України. – 2008. – № 9 : Інформатика та кібернетика. – С. 43-48.
2. Audio-visual answer to modern computing. – research*eu results supplement. - №26. – July/August 2010. – 31-32p.
3. Мічо Кайку. Фізика майбутнього / переклала з англ.. Анжела Кам'янець. – Львів: Літопис. – 2013. – 432.
4. Software: running commentary for smarter surveillance? .- Reasearch*eu results supplement. – №24. – May 2010. – P.29
5. V. Hrytsyk, A. Grondzal, A. Bilenkyj,. Augmented reality for people with disabilities // Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT'2015.
6. Validation of various adaptive threshold methods of segmentation applied to follicular lymphoma digital images stained with 3,3'-Diaminobenzidine&Haematoxylin. Published online 2013 Mar 25. Режим доступу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3656801/
7. Sauvola J, Pietikainen M. Adaptive document image binarization. Patt Recognit. 2000;33(2):225–236. doi: 10.1016/S0031-3203(99)00055-2. [] Sauvola J, Pietikainen M. Adaptive document image binarization. Patt Recognit. 2000;33(2):225–236. doi: 10.1016/S0031-3203(99)00055-2. [Robots of tomorrow with intelligent visual capabilities. research results magazine. - №62. – May 2017. – 38p.
8. Mehmet Sezgin, Bulent Sankur. Journal of Electronic Imaging 13(1), 146–165 (January 2004).- Survey over image thresholding techniques and quantitative performance evaluation
9. Zack, G.W., W.E. Rogers, and S.A. Latt, Automatic Measurement of Sister Chromatid Exchange Frequency. 1977. 25(7): p. 741-753.
10. Dronyuk, I.Development of printed packaging protection technology by means of background nets // Dronyuk, I., Nazarkevych, M. – Experience of Designing and Application of CAD Systems in Microelectronics - Proceedings of the 10th International Conference, CADSM 20092009, Article number 4839864, Pages 401-403.
11. Cutting-edge facial recognition goes mainstream. – Reasearch*eu results magazine. – №68. – December 2017-January 2018. – P.39

Author Index