THE EFFECTIVE WORKING WITH TOVEK TOOLS

Ladislav Burita
University of Defence and Tomas Bata University, Czech Republic
ladislav.burita@unob.cz

Kamil Halouzka
University of Defence, Czech Republic
kamil.halouzka@unob.cz

Abstract:
The article presents the experiences with using the software (SW) Tovek Tools (TT) in information analysis (IA) and in knowledge management systems (KMS) development. The special focus is oriented at the education in IA. The functions of the TT modules are described: Index Manager for indexing of source documents, Tovek Agent for information retrieval, Query Editor for complex query preparation, InfoRationg for documents context analysis, and Harvester for content analysis.; they can be used as a text mining tool. In the recherché of the literature are mentioned tools and methods of the IA that have similar functions as the TT, but TT is more complex. Chapters about using TT for KMS development and about education with TT are the core of the paper and summarized the author's experiences. The methodology approaches in ontology preparation for the KMS and steps in education with TT are mentioned.

Keywords: information analysis, Tovek Tools, knowledge management system, text mining

1. INTRODUCTION

Everyone need information in all possible appearances for its work and spend a lot of time processing it. Amount of information increases rapidly every year and people are looking for possibilities how to effectively find, understand and use information they want to work with. Of course, people want to process structured and unstructured data from different sources like local data or remote servers. It is important to speed up an operative work with information and to do it more effective. It means to explain people the easy way to find information accurately, quickly and intuitively from many sources. It is easy to find needed information but it is not easy to analyse large amount of information in short period and discover important associations and coherences. It is easy to analyse one article but it is a problem to analyse a huge amount of unstructured data.

One of the software (SW) tool than can be effectively use in information analysis is SW Tovek Tools (TT) that is successfully implemented in research and education at the University of Defence in Brno and Tomas Bata University in Zlín, Czech Republic.

The goal of the paper is to present the SW TT in information analysis (IA) for the knowledge management systems (KMS) development and in education. After introduction are described functions of the TT, they are specified modules Index Manager, Tovek Agent, Query Editor, InfoRationg, and Harvester. Next chapters about KMS and education with TT are the core of the paper and summarized the author's experiences.

2. FUNCTIONS OF THE TOVEK TOOLS

The TT is a desktop application designed for information searching, various types of analysis and the creating of summaries and reports. This is suitable for working with large amounts of text data from various informational sources using module for indexing of sources (Index Manager), for information retrieval (Tovek Agent), for context analysis (InfoRating), and content analysis (Harvester) and to produce recherché, summarizations, graphs, and reports (www.tovek.cz).
The TT is mostly used in news reporting, research and development in the investigation and detection of crime, or mapping the competitive environment. In the commercial sector are used mainly in media agencies, providers of information or in the financial sector, banks and insurance companies. In public administration they are mainly used in government and public institutions, the ministries, police forces and the security and intelligence community.

2.1 Components of the TT

Index Manager performs manual or automatic indexing various types of documents, database records (using ODBC-Open DataBase Connectivity) and e-mail messages. It uses fast and reliable filters that support all common data formats (MS Office, Open Office, PDF, etc.). The result of indexing is to create a new or update existing full-text resources that allow the user to quickly access information from a single environment without having to move data from their original locations. Index Manager also allows you to connect to existing full-text resources. The last version of TT is able to index sources in many languages: Czech, English, Arabic, Chinese, Danish, Finnish, French, Italian, Japanese, Korean, Hungarian, Norwegian, German, Polish, Portuguese, Rumanian, Russian, Greek, Slovenian, Slovak, Spanish, Swedish, Turkish, and Vietnamese.

Tovek Agent serves as the interface to search in indexed documents from the information sources through queries and the search results are displayed. Thanks to the features used technology offers advanced search methods, including evaluation of the relevance of retrieved documents and their alignment, automatic clustering of documents by the common content (clustering) search in different languages simultaneously and automatically create annotation of documents. To view the found documents it is possible different variants of outputs in many common formats (txt, html, xml).

Query Editor is intended for expert users, which helps to create and debug structured full-text queries (TOPIK) used for an accurate search in large volumes of data. TOPIK allows specifying in the form of a

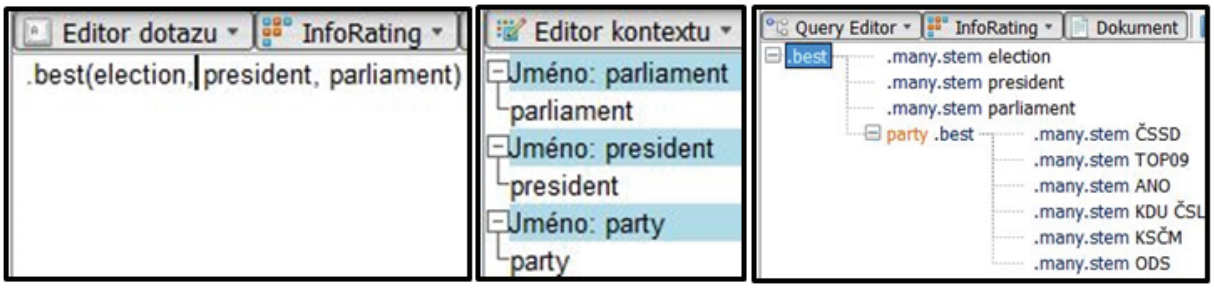
visual tree exactly all words, phrases and other search features, including their weights, which can be used to describe the search object.

InfoRating performs contextual analysis of selected documents, which finds and shows relations between the contents of found documents and defined themes. The result is shown in the form of matrix or in graph. The module can also be used to detect relationships between individual subjects or for creating structured searches by applying cross references.

Harvester allows performing content analysis of unstructured texts, especially for automated determination of the topic of the document; to select the documents with similar themes, and to identify new topics in the sets of documents (identification of trends). Topics are identified, based on statistical analysis of occurrence of close words pairs and triples. The connection variants can be presented graphically.

The TT is improved by a number of years; it is accessible in different versions, currently in version 7.3, in which the modules Tovek Agent, InfoRating, and Query Editor are integrated into a single user interface; see Picture 1.

Picture 1: Interface for Tovek Agent, InfoRating, and Query Editor



Source: authors

2.2 Search in Tovek Tools

The Tovek Agent is a module that is used for search in documents. A query can be written as:

1. Boolean expression.
2. Free text.
3. Topik - hierarchically assembled question.

Before user run the query, set the source for the search. Result of a query (HIT) is in synoptic tabular form and includes the selected documents that match a given criteria (see Picture 2, left part). The Query Editor module is used for the creation of so-called Topik which defines the terms of the search. The Topik contains words that are arranged in a hierarchical structure; Picture 1 (right part).
The text of the selected document, from the list, displays Tovek Viewer; found key words of the query are highlighted (see Picture 2, right part).

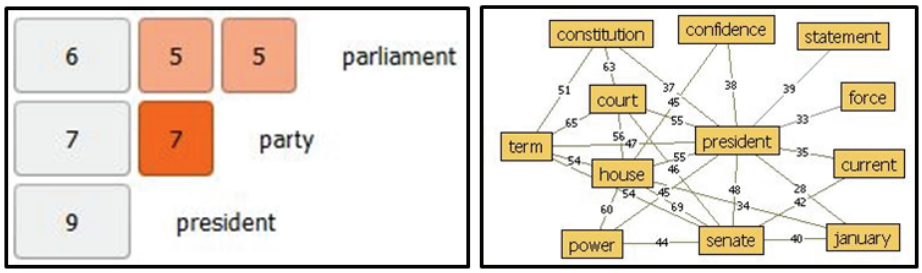
Picture 2: The result of TT question; document in TT viewer

Source: authors

2.3 The analytical features of the TT

The analytical features of the TT include modules InfoRating and Harvester. InfoRating module provides an overview of the context of the selected documents, based on contextual queries, see Picture 1 (central part). The results of contextual queries are relations between documents in the form of a chart or matrix, see Picture 3 (left). The content analysis of selected documents is performed in Harvester, which analyses the content of the input documents and arranges relevant words and their relationships. It uses a range of statistical methods to determine whether a word or a relationship between two words is relevant. Relevance depends on the distribution of the word in analysed documents, see Picture 3 (right). Harvester allows viewing of analysis results by typing relevant word; graphical browsing of relations between relevant words; examining clusters of relevant words; tracking the surrounding of relevant words; and finding trends.

Picture 3: Contextual matrix of InfoRating and the content analysis of Harvester



Source: authors

3. INFORMATION ANALYSIS IN THE LITERATURE

The literature recherché about information analysis (IA) is dealing with the two parts, the first is about tools for IA and the second is about methods of IA.

3.1 Tools for information analysis

There are many software (SW) products that are focused on information retrieval and IA that include indexing if information sources (documents, e-mails, databases) and searching in the indexed sources. Part of that SW product implements subsequent work with the sources such as establishing of patterns, determining of the relevance and implementation of context and content analysis.
One of the examples of that SW is Text Analytics from OdinText Company (OdinText, 2016) that involves applying advanced statistics and other machine learning techniques to text data in order to find patterns and discover important relationships, which leads to valuable insights. This analytics tools have a rules based approach. The problem with this SW is that its rules do not transfer well between industries, categories and various types of data, and therefore require customization.

The second example is DataMatch SW from Data Ladder Company (Data Ladder, 2016) that can be used to find and link customer data, consolidate data across multiple sources, and remove deceased and unwanted records - quickly and easily improving user marketing and mailing performance. Mentioned software main features:

- Semantic recognition of structured and unstructured product data
- Capability to cleanse, match, govern, and validate product data from any source
- Integrated governance capabilities for data stewards and product specialists
- Includes advanced classification and translation capabilities
- Advanced semantic matching capabilities

Data Ladder Company distributes also ProductMatch software that uses semantic and machine learning technologies to recognize and change difficult and complex product data from multiple sources.
The third example is Clustify text analytics software (Clustify, 2016), used in e-discovery and other fields. It can analyse millions of documents on a modest desktop computer. Documents can be stored in a database or as individual files on disk, allowing Clustify to work with many other platforms and tools. Clustify offer real-time predictive coding, immediately showing the impact of reviewing a training document. Clustify also offers document clustering. Clustering can group documents that are conceptually similar, near-duplicates, or part of an email thread.
Information analysis in the form of text mining is deeply developed in the article “Mapping the field of communication technology research in Asia: content analysis and text mining of SSCI journal articles 1995-2014” (Zheng & Liang & Huang & Liu, 2016). The article describes communication technologies in Asia's robust economic, cultural, and technological performance in the current century; this study maps the landscape of communication technology research in Asia of the recent two decades. Using a combination of content analysis and text mining-based semantic network analysis, this paper reviews 272 articles on Asian communication technology published in SSCI communication journals between 1995 and 2014.
The last example of article describing text mining is “Analysis of effectiveness of tsunami evacuation principles in the 2011 Great East Japan tsunami by using text mining” (Yun & Lee, 2016). This article analyses an effectiveness of tsunami evacuation principles from descriptive comments from the survivors and the non-survivors in the 2011 disaster using text mining method. As a result using the Naive Bayesian classifier, it identifies some of the evacuation behaviours differences taken by the survivors or by the non-survivors under the disaster as an effectiveness of the evacuation principles, and attempts to understand how to provide more practical instructions. These article results give effective recommendations for evacuation preparation against catastrophic earthquake and tsunami disasters in the future.

3.2 Methods and results of information analysis

Information analysis (IA) is performed from any source, such as text files, Web pages, e-mail, database, but also from recording audio or video. There are approaches that address IA in an integrated form from a variety of sources, such as TT.

IA of the web sources is mentioned in (Evrim & McLeod, 2014). Context-Based IA investigates the context information of the user's information request to provide relevant results for the given domain users. The relevance is measured by the semantics of the documents. The information extracted from lexical and domain ontologies are integrated by the user's interest information to expand the terms entered in the request. The obtained set of terms is categorized and the relations between the categories are obtained from the ontologies. This categorization is used to improve the quality of the document selection.

The scientific and technology information is an interest of the IA from the Lattes database (Silva & Smit, 2009). An exploratory study is presented which was developed with CVs with the goal to identify whether the open nature of the system could put at risk the consistency of the data when information is retrieved. Some suggestions were designed to improve the system from a perspective of results systemization. IA focuses on the analysis of the document content, on the choice of descriptors (Class Specification) given subject area (e.g. for preparing ontology). IA also investigates the context between documents (Evrim & McLeod, 2014).

The identification of molecular descriptors that contain compound class-specific information has a high relevance in chemo informatics (Wassermann & MaiNisius & Vogt & Bajorath, 2010). A generally

applicable way to identify such descriptors is to determine and compare their information content in a given compound activity class and in large databases where the vast majority of compounds do not have the desired activity. For this purpose, the Shannon entropy concept from information theory can in principle be employed. The methodology to reliably select descriptors by transforming the previously introduced differential Shannon entropy formalism was implemented into mutual information analysis.
Regarding the areas of IA, it is obviously that include science and research (Silva & Smit, 2009), but also the political aspects of science (Budd, 2007). In the paper are stated political and ideological claims about climate change are themselves reflected in the governmental and popular records. The examination further demonstrates that the governmental and popular records are informed not by scientific research and communication but by ideological stances.

Very important topic of IA is geographical information system (GIS) and geographic data in general (Devillers & Bedard & Jeansoulin & Moulin, 2007). The paper presents the design a tool that can manage heterogeneous data quality information and provide functions to support expert users in the assessment of the fitness for use of a given dataset. Combining concepts from GIS and Business Intelligence, this approach provides interactive, multi-granularity and context-sensitive spatial data quality indicators that help experts to build and justify their opinions.

Typical areas of IA are literary sources (Eirao & da Cunha, 2013). It presents the results from searches in three sources of information: Library Information Science Abstract, Annual Review of Information Science and Technology and Journal of Documentation, about selective dissemination of information and RSS technology in libraries. The results demonstrated that there is a significant number of articles published in several languages and journal articles about the themes, especially in 70's and after a period of decrease of publications.

The last information concerns IA in commercial sphere, which is offered as a service (Hacigumus & Rhodes & Spangle & Kreulen, 2006). The paper presents the architecture of a Business Information Analysis provisioning system, BISON. The service provisioning system combines two prominent domains, namely structured/unstructured data analysis and service-oriented computing.

4. KNOWLEDGE MANAGEMENT USING THE TT

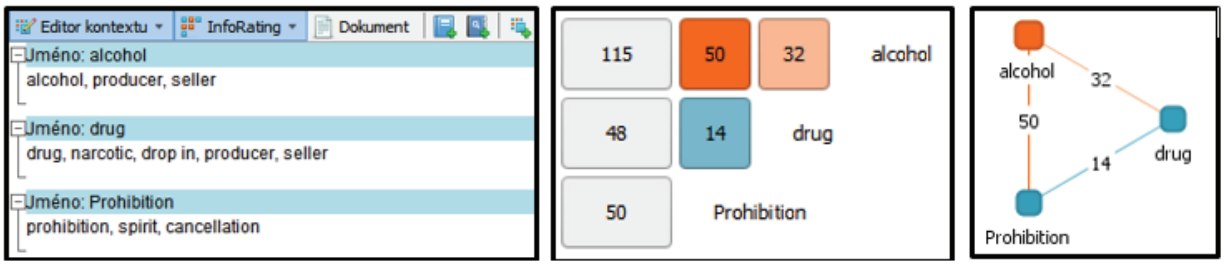
The chapter will describe a practice use of TT in knowledge management (KM). The SW TT is suitable for processing even large volumes of unstructured data from miscellaneous information sources. It is possible to use local data sources or to connect to Tovek Server, where are available information on demand of user. For the KM goals is TT used in early stage of KMS development. User can map out environment that is interested in. Data sources will be used for searching and analysing of themes, basic concepts as candidates for ontology classes. For example, it is possible to look for alcohol, drug and prohibition and TT find all relevant documents concerning that query. A full text query can consist of several sub-queries in different languages which are used on documents in the relevant language.

The search results are combined into one common output for further work. Directly is possible to require weights for each query and directly specify what is important. For better further analysis can user labelled founded retrieved documents as favourite. The search results can be then analysed using context and content analyses, when TT offer and user choose founded entities as person, state, date, currency and city. In that case, can user very simply find connection between various kinds of crime. It means that one person or company can distribute alcohol in defined towns and date. The search results and entities are clearly and colourfully highlighted. The user can thus easily discover relationships or identify new trends and topics. User can chose set of documents and export them in several common formats (txt, html, xml). One of the next steps of IA is context analysis that is suitable for finding relationships between documents. User can define themes, for example:

- Alcohol – alcohol, producer, seller.
- Drug – drug, narcotic, drop in, producer, seller.
- Prohibition – prohibition, spirit, cancellation.

At the end, TT show contextual matrix with relationships between retrieved document contents and defined themes, see Picture 4. The particular contextual questions identify 115 occurrences of theme alcohol, 48 occurrences of drug and 50 occurrences of prohibition. The context between alcohol and prohibition was found in 50 documents, between alcohol and drug was found in 32 documents, and between drug and prohibition in 14 documents.

Picture 4: The InfoRating contextual information in theme “alcohol-drug-prohibition”



Source: authors

The TT is able to excellently work with unstructured documents and can automatically identified topics (for example alcohol) and connection among them, can found topics words neighbourhood and can identify document matching to the selected topics. User can for example identify neighbourhood of “***word***” alcohol (in Czech “***slovo***”), see Picture 5.

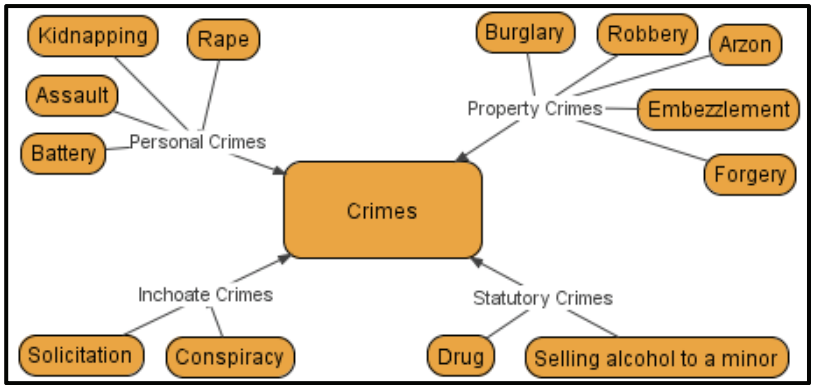
Picture 5: The topics words neighbourhood

4< Slovo	3< Slovo	2< Slovo	1< Slovo	Slovo	Slovo >1	Slovo >2	Slovo >3	Slovo >4
3ae7261c-5d07-4f68-b58f-cb65373d5302								
crime	government	health	drink	alcohol	title	czech	ministry	plan
czech	ministry	plan	ban	alcohol	export	eu	may	ban
czech	ministry	plan	ban	alcohol	export	eu	may	ban
deputy	brussels	might	ban	alcohol	import	czech	republic	government

Source: authors

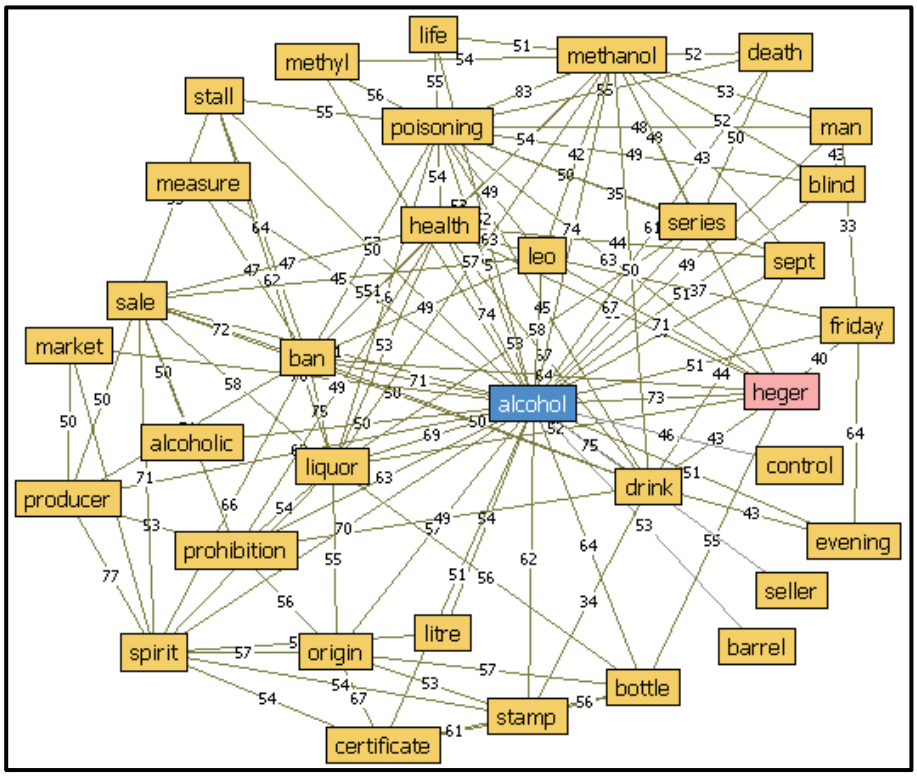
The main goal in the KMS preparation using TT is an excellent understanding of the source documents, the taxonomy preparation (the simple knowledge form of the subject area); see Picture 6. Next searching for the candidate for ontology classes; see Picture 7. At the pictures are generally depicted types of crimes that can be divided into four major categories: personal crimes, property crimes, inchoate crimes, and statutory crimes.

Picture 6: Taxonomy of the subject area



Source: authors

Picture 7: New classes identification



Source: authors

5. EDUCATION USING THE TT

Teaching using TT focuses on proper understanding of how to perform an IA. The education about the TT precedes lectures in working with text and Document-IS (DIS). In particular, emphasis is placed on Boolean search and document processing stages in the DIS. Followed by familiarization with the TT modules in practical work. Teacher explains and demonstrates the work with modules and their function in IA. As an ideal data source can serve demo data of TT, focused on problems of political and social life, so the IA is relatively easy. Strategies for document search and analysis, of the obtained results, should be derived from the principles of competitive intelligence, see Picture 6.

Picture 6: Competitive intelligence objects and procedures

Source: authors

Education continues in assignment tasks for independent work. Specified topics for students work in IA are constantly changing, with respect to the ability of students and the goals of the research work of the department, but also due to the development of students' skills to work independently and creatively solve the challenges.

Tasks are usually described only briefly. A set of documents for IA either selects the teacher or the students seek themselves. The range is some dozens of documents. Among the specified topics have been included the files of WikiLeaks, documents from military environment (weapons, foreign missions, terrorism, NATO) or social issues (unemployment, elections, Islamic state, the situation in Ukraine, university education) or technology topics (energy, mobile applications, SMART technology, the future of the internet, cyber security).

The students task is a credit work, in which must be described the strategy of IA, must be used all modules of TT, and must finished into overall conclusions with findings of the new ideas.

6. DISCUSSION

The SW TT is closely connected with information analysis and knowledge management system. This SW helps university students to understand the work with structured and unstructured data from various information sources and shows ways of possible use in research and development. The various types of analyses are very useful procedures that facilitate students to find and understand logical information associations in the form of content and context analysis.

Students greatly appreciate the opportunity to view the results in text and graphical form. University of Defence team has great practical experiences of TT in designing of classified information security analysis where is the main attention focused on personal security, physical security, administrative security, cryptographic security, and communication and information system security. The next team work could be focused on IA using the text mining methods.

7. CONCLUSIONS

Paper summarizes the authors' experiences with TT for research and education. The literature recherché shows some topics and goals for information analysis and tools used in that activity. The described SW TT is an excellent tool, in comparison with the others, for information retrieval; for content and contextual analysis. There are explained functions of the TT modules and are mentioned some examples of the effective use in education and research.

The research application of the TT is mostly used in KMS development at the earlier stages for analysis of information sources, for good understanding of the research focus, and for the ontology design. The methodology steps are in the paper illustrated. Education using the TT makes possible to introduce elements of the individual, creative and independent work.

8. ACKNOWLEDGEMENT

The article presents the results and experiences of the research and education using the SW TT. Its application in university environment is connected with the project (ZRO209, 2016) that follows in the analysis, development, and implementation KMS and using them in Army of the Czech Republic and teaching at the University of Defence, Faculty of Military Technology, Department of Communication and Information Systems. The next source of the experiences is the Tomas Bata University in Zlín, Faculty of Management and Economics, Department of Industrial Engineering and Information Systems.

REFERENCE LIST

1. Budd, J. M. (2007). Information, analysis, and ideology: A case study of science and the public interest. *Journal of the American society for information science and technology, 58 (14),* 2366-2371. DOI: 10.1002/asi.20703.
2. Clustify (2016). Retrieved from: http://www.cluster-text.com/.
3. Data Ladder (2016). Retrieved from: https://dataladder.com/.
4. Devillers, R., & Bedard, Y., & Jeansoulin, R., & Moulin, B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International journal of geographical information science, 21(3)*, 261-282. DOI: 10.1080/ 13658810600911879.
5. Eirao, TG., & da Cunha, MB. (2013). Selective dissemination of information: analysis of the literature published during 1958-2012. *Informacao & sociedade-estudos, 23(1),* 39-47.
6. Evrim, V., & McLeod, D. (2014). Context-based information analysis for the Web environment. DOI: 10.1007/s10115-012-0493-x.
7. Hacigumus, H., & Rhodes, J., & Spangler, S., & Kreulen, A. (2006). BISON: Providing Business Information Analysis as a service. In *Advances in database technology - EDBT 2006* (pp. 1084-1087). Berlin, Germany: Springer-Verlag.
8. MilUNI (2014). A portal for military universities cooperation. Brno, Czech Republic: University of Defence. Retrived from http://miluni.eu

9. Odintext (2016). Retrieved from: http://odintext.com/about-odintext/.
10. Silva, FM., & Smit, JW. (2009). Information organization in open electronic systems of scientific and technological information: analysis of the lattes database. *Perspectivas em ciencia da informacao, 14(1)*, 77-98.
11. Wassermann, A., & MaiNisius, B., & Vogt, M., & Bajorath, J. (2010). Identification of Descriptors Capturing Compound Class-Specific Features by Mutual Information Analysis. *Journal of chemical information and modeling, 50(11*), 1935-1940. DOI: 10.1021/ci100319n.
12. Yun, NY., Lee, SW. (2016). Analysis of effectiveness of tsunami evacuation principles in the 2011 Great East Japan tsunami by using text mining, *Multimedia tools and applications*, 75(20), 12955-12966. DOI: 10.1007/s11042-014-2326-2.
13. Zheng, P., Liang, X., Huang, GX., Liu, X. (2016). Mapping the field of communication technology research in Asia: content analysis and text mining of SSCI journal articles 1995-2014. *Asian Journal Of Communication*, 26(6), 511-531. DOI: 10.1080/01292986.2016. 1231210.
14. ZRO209 (2016). Project for development of organization. Brno, Czech Republic: University of Defence, 2016-2020.