ECONTECHMOD. AN INTERNATIONAL QUARTERLY JOURNAL – 2017. Vol. 6. No. 2. 73–78

Analysis of methods and means of text mining

Z. Rybchak [1]*, O. Basystiuk* [2]

[1]Lviv Polytechnic National University, e-mail: zoriana.rybchak@gmail.com
[2]Lviv Polytechnic National University, e-mail: obasystiuk@gmail.com

Received February 15.2017: accepted May 28.2017

Abstract. In Big Data era when data volume doubled every year analyzing of all this data become really complicated task, so in this case text mining systems, techniques and tools become main instrument of analyzing tones and tones of information, selecting that information that suit the best for your needs and just help save your time for more interesting thing. The main aims of this article are explain basic principles of this field and overview some interesting technologies that nowadays are widely used in text mining.

Key words: text mining, text analytics, data analysing, high-quality information, text categorization, text clustering, document summarization, sentiment analysis.

INTRODUCTION

Text mining, also known as intelligent text analysis, text data mining or knowledge-discovery in text (KDT) all this terms describes a set of linguistic, statistical, and machine learning techniques that refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Techniques that help model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation.

Text mining is a relatively young field, manual text mining researches started in the middle of 1980s, notably for sciences and government needs, but interested people and technological advances have enabled developing of that field during this time. Nowadays this field united such field of science, like: computational linguistics, machine learning and some specific field of statistics.

The field of text mining usually deals with text whose main function is providing communication and help people to express their thoughts and opinions, and the motivation for trying to extract information from such text automatically is compelling – even if success is only partial.

The term text mining also describes that application of text mining to respond to business problems, whether independently or in conjunction with query and analysis of fielded, numerical data. It is a truism that more than 70 % of business-relevant information is stored in unstructured form, such as text. These techniques and processes help to discover and present knowledge – facts, business rules, and relationships – that is otherwise locked in textual form, impenetrable to automated processing.

As said earlier 70% of business-relevant information is stored as text, but this is truism for all spheres of human life, most information is currently stored as text information, that why text mining is believed to have a high commercial potential value. Increasing interest is to multilingual data mining: the ability to gain information across languages and cluster similar items from different linguistic sources according to their meaning.

A simple application is to scan a text written in a natural language, then identify key-phrases of this document, show this key-phrases and by this phrases program can make prediction of what type of text it is.

Typical tasks in text mining include this: text categorization, and text clustering (this tasks is main goal of my example of simple application), also it could be concept extraction, entity extraction, sentiment analysis (popular, when you want to know, what people think about your business), document summarization etc.

This technology nowadays is widely implemented and used in variety of government, research, and business needs.

TEXT MINING AND MINING TECHNICS OF RECEIVING HIGH-QUALITY INFORMATION

Text mining, it`s process of mining text data, or in other words receiving high-quality information from text data. High-quality text information usually refers to some combination of relevance, novelty, and interestingness. To get high-quality information from text, use few methods, such as: information retrieval, lexical analysis to study word frequency distributions, pattern recognition, pattern learning, regularities in data, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. In addition, text mining involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output.

The main goal is, to turn text into data for analysis, by applications of natural language processing and analytical methods, and when you get this data for analysis, you can create method of processing this data in way you need, or you already created libraries for this.

All this method is a branch of machine learning (or nearly synonymous with machine learning), especially recognition of patterns and regularities in data, because pattern recognition systems are in many cases trained from labeled “training" data, but when no labeled data are available other algorithms can be used to discover previously unknown patterns. Same thing for regularities, in this case knowledge of computational linguistics is needed, you input labeled data of regularities and trained your algorithm to recognize this regularities and handle them in right order.

TECHNIQUES AND TOOLS ANALYSIS

Text mining systems use a big spectrum of different approaches, partly because of the great scope of systems and tools that perform text mining, and partly because the field don`t have it dominant methodologies due to the youth of this field. Nevertheless, we can divide these approaches on:

- High-level, that’s mean you involved into your text mining application systems that use an automatic training systems to do stuff like recognising patterns. This works in easy way, you create some training system, push into this system already target “training” data and by this data your system learn how to recognize different patterns.
- Low-level, that`s mean you deal with natural language and involve your custom created decisions making system, that strongly influence the success of mining process.

But as low-level way of solving problems is more complexity, time-consuming and required big amount of knowledge, it`s less popular nowadays. Because you need to implement by yourself this trivial on first view, but really important logic on dealing with all small decisions, like: how to deal with apostrophes and hyphens, capitalization, punctuation, numbers, alphanumeric strings, whether the amount of white space is significant, whether to impose a maximum length on tokens, what to do with non-printing characters, and so on. So using of this method require good background in text analysis principles and implementing all logic of this small, but really important, decisions.

In this case, high-level way provides big opportunities to create some small, but pretty useful app in a few minutes. There are a lot of text mining computer programs available from many commercial and open source companies and sources. Most popular of them are:

Commercial:

- ABBYY Compreno;
- Mathematica (provides built in tools for text alignment);
- Luminoso;
- IBM LanguageWare and IBM SPSS.

Open source:

- Natural Language Toolkit;
- OpenNLP;
- Orange.

One particular framework and development environment for text mining, called General Architecture for Text Engineering or GATE, aims to help users develop, evaluate and deploy systems for what the authors term “language engineering.” It provides support not just for standard text mining applications such as information extraction, but also for tasks such as building and annotating corpora, and evaluating the applications.

At the lowest level, GATE supports a variety of formats including XML, RTF, HTML, SGML, email and plain text, converting them into a single unified model that also supports annotation. There are three storage mechanisms:

- relational database;
- serialized Java object;
- XML based internal format documents can be re-exported into their original format with or without annotations.

Text encoding is based on Unicode to provide support for multilingual data processing, so that systems developed with GATE can be ported to new languages with no additional overhead apart from the development of the resources needed for the specific language. GATE includes a tokenizer and a sentence splitter. It incorporates a part-of-speech tagger and a gazetteer that includes lists of cities, organizations, days of the week, etc. It has a semantic tagger that applies hand-crafted rules written in a language in which patterns can be described and annotations created as a result.

Patterns can be specified by giving a particular text string, or annotations that have previously been created by modules such as the tokenizer, gazetteer, or document format analysis. It also includes semantic modules that recognize relations between entities and detect co-reference. It contains tools for creating new language resources, and for evaluating the performance of text mining systems developed with GATE.

One application of GATE is a system for entity extraction of names that is capable of processing texts from widely different domains and genres. This has been used to perform recognition and tracking tasks of named, nominal and pronominal entities in several types of text.Least but not last, there are few popular techniques, which will help you to start work in effectively mining text data.

- Sentiment analysis – analyzing the people opinions and tone of feedback, posts, articles about company in social media, also involved researching in call center information. This technique help business to collect information about their products or services faster. And based on this information see how products or services is performing in the market, find out what customers are saying about competitors, and so on. Finally, use all this information for improving your business for better meet people's needs.

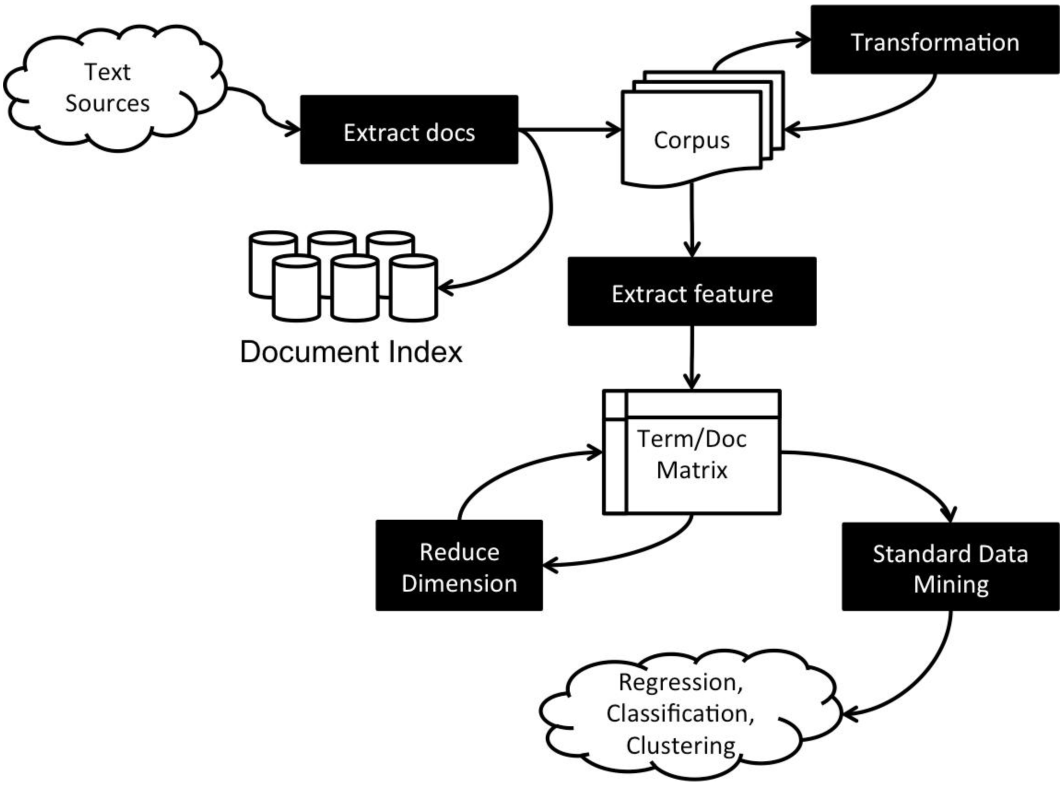

Fig. 1. Scheme of most popular text mining tools

- Topic modeling – useful technique for identifying main theme of document. Legal firms, use this technique for going through millions of documents in big litigation cases.
- Term frequency – inverse document frequency – looks at how frequently a word appears in a document and its importance relative to the whole set of documents.
- Named entity recognition – this technique used for recognizing nouns. In this case, program could be used for fetching persons or organizations names, geographic locations, monetary amounts, dates from text information. This technique analyzes surrounding words and by this recognize the nouns.
- Event extraction – this is more complex technique than named entity recognition and harder to organize, because it is not only looks at nouns and it surrounding words, but also recognize the relationships between them. That's mean that we can analyze our text more complex and fetch information like: what is the main idea of this sentence, how this sentence was organized, etc.

TEXT MINING VS INFORMATION EXTRACTION

Information Extraction (IE) is the process of automatic extraction of structured information such as entities, relationship between entities and attributes describing entities from unstructured texts. Mostly useful information such as names of people, places or organization mentioned in the text is extracted without a proper understanding of the text. Traditional data mining systems assumes that the information to be mined is already in the form of relational database [17].

Information Retrieval deals with the problem of finding relevant documents in a collection. Information Extraction identifies useful relevant text in a document. Useful information is defined as text segment and its associated attributes.

TEXT MINING VS INFORMATION RETRIEVAL

Information Retrieval (IR) is finding a document of an unstructured nature usually text that satisfies an information need from within large collections usually stored on computers. Information retrieval is fast becoming the dominant form of information access, overtaking traditional database style searching and IR can also cover other kinds of data and information problems beyond that specified in the core definition above.

These range from fully linguistic (based on parsing the sentences) to fully statistical (e.g., based on counting word co-occurrences). While doing an IR research it is proved that phrases are valuable indexing units and yield improved search effectiveness however, the style of phrase generation used is not that critical. Studies comparing linguistic phrases to statistical phrases have failed to show a difference in their retrieval performance.

Some IR systems also use multi-word phrases (information retrieval) as index terms. Since phrases are considered more meaningful than individual words, a phrase match in the document is considered more informative than single word matches. Several techniques to generate a list of phrases have been explored. Automatic extraction of metadata is an important application of TM techniques. However, existing automatic document retrieval techniques bypass the metadata creation stage and work on the full text of the documents directly [Salton and McGill, 1983] in which the basic idea is to index every individual word in the document collection.

The documents are represented as a “bag of words” that is, the set of words that they contain, along with a count of how often each one appears in the document which makes it easier for retrieval. There are various IR systems available in which two systems AMORE and MAPBOT has been listed with its working mode as follows:

AMORE: Advanced Multimedia Oriented Retrieval

Engine [7] The Harvest Information Discovery and Access System for text indexing and searching, and using the content oriented image retrieval (COIR) library for image retrieval.

Advantage: Its an automatic indexing of both text and image from one or more Web sites.

MAPBOT: An interactive Web based map information retrieval system [11] in which Web users can easily and efficiently search geographical information with the assistance of a user interface agent (UIA). Each kind of map feature such as a building or a motorway works as an agent called a Maplet.

MAPBOT, an active map system using software agent technology is presented to solve these problems.

Advantage: Maplets to communicate with information agents outside the system to retrieve more information for the user.

There are many models available for IR process which can be broadly classified as:

- Classical models of IR based on mathematical knowledge that was easily recognized and well
- understood simple, efficient and easy to implement.

The three classical information retrieval models are:

Boolean, Vector and Probabilistic models.

- Non-Classical models of IR are based on principles other than similarity, probability, Boolean operations, etc. on which classical retrieval models are based on information logic model, situation theory model and Interaction model.
- Alternative models of IR. .Alternative models are enhancements of classical models making use of specific techniques from other fields. Example:” Cluster model, fuzzy model and latent semantic indexing (LSI) models.”

Generally, Patent documents contain important research results that are valuable to the industry, business, law, and policy-making communities. When carefully analysed, they might show technological details and relations, reveal business trends, inspire novel industrial solutions, or help make investment policy. Patent analysis or mapping requires considerable effort and expertise. As can been seen, these processes require the analysts to have a certain degree of expertise in information retrieval, domain-specific technologies, and business intelligence.

Now, let’s overview some field where text mining already is highly used and investigate what text mining applications are doing.

APPLICATIONS

In this section we briefly discuss successful applications of text mining methods in quite diverse areas as patent analysis, text classification in news agencies, bioinformatics and spam filtering. Each of the applications has specific characteristics that had to be considered while selecting appropriate text mining methods.

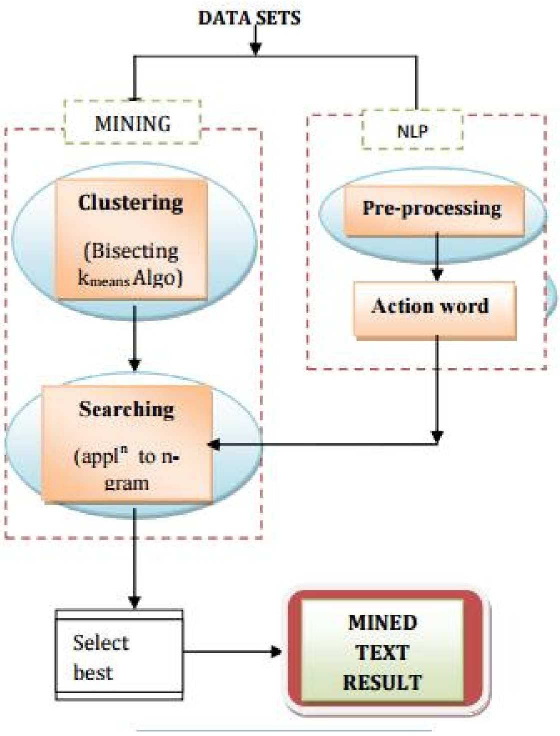

Fig. 2. Common architecture of text mining system

Patent Analysis. In recent years the analysis of patents developed to a large application area. The reasons for this are on the one hand the increased number of patent applications and on the other hand the progress that had been made in text classification, which allows to use these techniques in this due to the commercial impact quite sensitive area. Meanwhile, supervised and unsupervised techniques are applied to analyze patent documents and 27 to support companies and also the European patent office in their work.

The challenges in patent analysis consists of the length of the documents, which are larger then documents usually used in text classification, and the large number of available documents in a corpus. Usually every document consist of 5000 words in average. More than 140000 documents have to be handled by the European patent office (EPO) per year. They are processed by 2500 patent examiners in three locations. In several studies the classification quality of state-of-the-art methods was analysed. Classifying patent applications with winnow reported very good result with an 3% error rate for 16000 full text documents to be classified in 16 classes (mono-classification) and a 6% error rate in the same setting for abstracts only by using the Winnow and the Rocchio algorithm. These results are possible due to the large amount of available training documents. Good results are also reported in Automatic categorisation applications at the European patent office for an internal EPO text classification application with a precision of 81 % and an recall of 78 %.

Text clustering techniques for patent analysis are often applied to support the analysis of patents in large companies by structuring and visualizing the investigated corpus. Thus, these methods find their way in a lot of commercial products but are still also of interest for research, since there is still a need for improved performance. Companies like IBM offer products to support the analysis of patent text documents. Dorre describes in the IBM Intelligent Miner for text in a scenario applied to patent text and compares it also to data mining and text mining. Coupet does not only apply clustering but also gives some nice visualization. A similar scenario on the basis of SOM is given in intelligent patent analysis through the use of a neural network (J.-C. Lamirel, S. Al Shehabi, M. Hoffmann, and C. Francois. In ACL-2003 Workshop on Patent Corpus Processing, 2003).

Text Classification for News Agencies. In publishing houses a large number of news stories arrive each day. The users like to have these stories tagged with categories and the names of important persons, organizations and places. To automate this process the Deutsche Presse-Agentur (dpa) and a group of leading German broadcasters (PAN) wanted to select a commercial text classification system to support the annotation of news articles. Seven systems were tested with a two given test corpora of about half a million news stories and different categorical hierarchies of about 800 and 2300 categories.

Due to confidentiality the results can be published only in anonymized form. For the corpus with 2300 categories the best system achieved at an F1-value of 39 %, while for the corpus with 800 categories an F1-value of 79 % was reached. In the latter case a partially automatic assignment based on the reliability score was possible for about half the documents, while otherwise the systems could only deliver proposals for human categorizers. Especially good are the results for recovering persons and geographic locations with about 80 % F1-value.

In general there were great variations between the performances of the systems. In a usability experiment with human annotators the formal evaluation results were confirmed leading to faster and more consistent annotation. It turned out, that with respect to categories the human annotators exhibit a relative large disagreement and a lower consistency than text mining systems. Hence the support of human annotators by text mining systems offer more consistent annotations in addition to faster annotation. 28 The Deutsche Presse-Agentur now is routinely using a text mining system in its news production workflow.

Bioinformatics. Bio-entity recognition aims to identify and classify technical terms in the domain of molecular biology that correspond to instances of concepts that are of interest to biologists. Examples of such entities include the names of proteins, genes and their locations of activity such as cells or organism names. Entity recognition is becoming increasingly important with the massive increase in reported results due to high throughput experimental methods. It can be used in several higher level information access tasks such as relation extraction, summarization and question answering. Recently the GENIA corpus was provided as a benchmark data set to compare different entity extraction approaches. It contains 2000 abstracts from the MEDLINE database which were hand annotated with 36 types of biological entities. The following sentence is an example: “We have shown that interleukin-1 and IL-2 control IL-2 receptor alpha gene transcription in CD4-CD8 murine T-lymphocyte precursors”. In the 2004 evaluation four types of extraction models were used: Support Vector Machines (SVMs), Hidden Markov Models (HMMs), Conditional Random Fields (CRFs) and the related Maximum Entropy Markov Models (MEMMs). Varying types of input features were employed: lexical features (words), n-grams, orthographic information, word lists, part-of-speech tags, noun phrase tags, etc. The evaluation shows that the best five systems yield an F1-value of about 70 %. They use SVMs in combination with Markov models (72.6 %), MEMMs (70.1 %), CRFs (69.8 %), CRFs together with SVMs (66.3 %), and HMMs (64.8 %). For practical applications the current accuracy levels are not yet satisfactory and research currently aims at including a sophisticated mix of external resources such as keyword lists and ontologies which provide terminological resources.

Anti-Spam Filtering of Emails. The explosive growth of unsolicited e-mail, more commonly known as spam, over the last years has been undermining constantly the usability of e-mail. One solution is offered by anti-spam filters. Most commercially available filters use black-lists and hand-crafted rules. On the other hand, the success of machine learning methods in text classification offers the possibility to arrive at anti-spam filters that quickly may be adapted to new types of spam. There is a growing number of learning spam filters mostly using naive Bayes classifiers

CONCLUSION

Text mining is an interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics/ The text analysis is important technology of nowadays which will is helping people to work more productive and save time making your routine work by itself. In most cases, it is difficult to provide general and meaningful evaluations because the task is highly sensitive to the particular text under consideration. Generally accepted characterizations of what it covers do not yet exist. When the term is broadly interpreted, many different problems and techniques come under its ambit. In most cases it is difficult to provide general and meaningful evaluations because the task is highly sensitive to the particular text under consideration.

Document classification, topic modelling, entity extraction, pattern recognition, sentiment analysis, and filling templates that correspond to given relationships between entities, are all central text mining operations that have been extensively studied. Using structured data such as Web pages rather than plain text as the input opens up new possibilities for extracting information from individual pages and large networks of pages.

The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process.

REFERENCES

1. **Sholom M. Weiss, Nitin Indurkhya, Tong Zhang. 2015.** Fundamentals of Predictive Text Mining. Springer London Publishing House, 239 p.
2. **Emma Tonkin, Gregory J. L. Tourte, Stephanie Taylor. 2016.** Working with Text: Tools, Techniques and Approaches for Text Mining. Elsevier Science & Technology, 344 p.
3. **Sonali Vijay, Archana Chaugule. Pramod Patil. 2014.** Text Mining Methods and Techniques. International Journal of Computer Applications (0975–8887), Vol. 85, No. 17, p. 34.
4. **Rebecca Merrett. 2015.** 5 tools and techniques for text analytics. Available online at: http://www.cio.com.au/article/575209/5-tools-techniques-text-analytics/
5. **Daniel Gutierrez. 2015.** Text Analytics: The Next Generation of Big Data. Available online at: http://www.predictiveanalyticsworld.com/patimes/text-analytics-the-next-generation-of-big-data-061215/5529/
6. **ChengXiang Zhai, Sean Massung. 2016.** Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. Morgan & Claypool, 530 p.
7. **Vidhya. K. A, G. Aghila. 2010.** Text Mining Process, Techniques and Tools: an Overview. International Journal of Information Technology and Knowledge Management July-December 2010, Vol. 2, No. 2, pp. 613–622.
8. **Markus Hofmann, Andrew Chisholm. 2016.** Text Mining, Web Mining, and Visualization Use Cases Using Open Source Tools. Taylor & Francis, 500 p.
9. **Ashish Kumar, Avinash Paul. 2016.** Mastering Text Mining with R. Packt Publishing, Limited, 258 p.
10. **Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, Fred Damerau. 2010.** Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer Science & Business Media, 237 p.
11. **Ronen Feldman, James Sange. 2013.** The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 410 p.
12. **Tom Reamy. 2016.** Deep Text: Using Text Analytics to Conquer Information Overload, Get Real Value from Social Media, and Add Bigger Text to Big Data. Today Inc, 415 p.
13. **Sholom M. Weiss, Nitin Indurkhya, Tong Zhang. 2015.** Fundamentals of Predictive Text Mining. Springer London Publishing House, 239 p.
14. **Andreas Hotho, Andreas Nurnberger. Gerhard Paaß, 2015.** A Brief Survey of Text Mining.
15. **Anne Kao, Steve R. Poteet. 2009**. Natural Language Processing and Text Mining. Springer London Publishing House, 265 p.
16. **Danny Sullivan.** What Is Search Engine Spam? The Video Edition. Available online at: http://searchengineland.com/what-is-search-engine-spam-the-video-edition-15202
17. **Rasim M. Alguliev, Ramiz M. Aliguliyev, and Saadat A. Nazirova. 2011.** Classification of Textual E-Mail Spam Using Data Mining Techniques. Applied Computational Intelligence and Soft Computing, Vol. 2011, 8 p.
18. **Jiexun Li, Harry Jiannan Wang, Zhu Zhang and J. Leon Zhao. 2010.** “A Policy-based Process Mining Framework: Mining Business Policy Texts for Discovering Process Models”’ Decision Support Systems table of contents archive, Vol. 48, Issue 1, pp. 267–288.
19. **Ming Zhao, Jianli Wang and Guanjun Fan. 2008.** “Research on Application of Improved Text Cluster Algorithm in Intelligent QA System”, Proceedings of the Second International Conference on Genetic and Evolutionary Computing, IEEE Computer Society, pp. 463–466.
20. **Ulianovskaya, Yu. 2016.** Information technology for treatment of results expert estimation with fuzzy character input data. ECONTECHMOD. An International Quarterly Journal, Vol. 5, No. 3, pp. 55–60.