SEMANTIC ANALYSIS ALGORITHMS FOR KNOWLEDGE WORKERS SUPPORT

Nina RIZUN[1], Mariia RIZUN[2], Yurii TARANENKO[3]

[1] Gdansk University of Technology, Poland,
nina.rizun@zie.pg.gda.pl
[2] University of Economics in Katowice, Poland,
mariia.rizun@uekat.pl
[3] Alfred Nobel University, Dnipro, Ukraine,
taranen@rambler.ru

Abstract: The paper examines various aspects of text analysis application for knowledge worker's activity realization. Conclusions are drawn about the relevance and importance of processing the non-structured textual information in order to increase knowledge worker's efficiency, as well as their awareness in different branches of science. The paper considers the existing algorithms of texts semantic analysis as the sphere of documents topical closeness recognition. At the same time, it contains an example of applying the complex methodology of semantic analysis, which includes LSA and LDA methods together with the Zipf's Law with the objective to solve a typical knowledge worker's task. Quantitative identifiers of the efficiency of this methodology are given.

Key words: Knowledge Worker, Latent Semantic Analysis, Latent Dirichlet Allocation, Zipf's Law.

1. Introduction

In today's economy knowledge creation and distribution are becoming more and more crucial factors of any organization's competitiveness. Knowledge is being thought of as a valuable commodity that is embedded in products and in the tacit knowledge of highly mobile employees. Because of that, the ability to manage knowledge is currently supposed to be the key ability for organizations. Knowledge management (KM) represents a deliberate and systematic approach to ensure the full utilization of the organization's knowledge base, coupled with the potential of individual skills, competencies, thoughts, innovations, and ideas to create a more efficient and effective organization (Rizun, 2017a).

In the study of knowledge management phenomenon, it is necessary to consider two elements, which are vital for the KM process: the people, who create, utilize and distribute knowledge – knowledge workers; and the technological tools (solutions), which assist knowledge workers and facilitate their activity (Rizun, 2017b).

According to Liao (2003), knowledge management technologies as tools for knowledge capturing and creation can be represented by 7 major categories: KM Framework; Knowledge-Based Systems; Data Mining; Information and Communication Technology; Artificial Intelligence / Expert Systems; Database Technology; Modeling.

And one of the most important techniques of the Artificial Intelligence/Expert Systems as well as Data Mining are the text-mining, text summarization and sentiment analysis.

In accordance with (Silwattananusarn and Tuamsuk, 2012), the problems of text-mining and natural language processing can be observed almost in any sphere of knowledge workers' activity. Table 1 contains examples of text mining application in the major branches of economics.

Table 1

Text Mining Application in the Major Branches of Economics

Sphere of knowledge work	Text mining examples
Health Care Organization	Textual data mining (literature, admission notes, reports, summaries). Discovering unsuspected links from the huge amount of literature and supporting medical research. Exploration of diseases, operations, and tumors relationships. Adding meaning to data semantic metadata (Hwang et al., 2008; Ghattas et al., 2010)
Retailing	Analysis of the market knowledge of customers, brands, products. Recognition of the sentiment of customers' opinions. Prediction of customers and clients behavior (Liao et al., 2008; Bose and van der Aalst, 2009)
Financial/Banking	Recognition the sentiment of the reputation of customers, banks and companies. Classification and quick retrieval of documents (Cheng and Sheu, 2009; Silwattananusarn and Tuamsuk, 2012)
Small and Middle Businesses (food company and food supply chain)	Help with information overload and overlook. Integration of knowledge from many sources. Enhancing decision support systems (Li et al., 2010a; Li et al., 2010b)
Entrepreneurial Science	Specific textual knowledge extraction. Help for managers in transformation of tacit knowledge to the explicit (Cantú and Ceballos, 2010)
Business	Discovering hidden patterns between textual data (reports, articles, reviews). Integration of knowledge from many sources. Enhancing decision support systems (Wu et al., 2010).
Collaboration and Teamwork	Analysis of the documents, e-mails, notes for revealing the common opinions and construction of worker's explicit knowledge flow. Mining and construction of the group-based tacit knowledge prototype for explicit data bases creation (Liu and Lai, 2011)
Construction Industry	Handling textual information source for industrial knowledge discovery and management solutions (Wang and Fan, 2008)
Education	Classification of textual data sources (literature, articles, administrative papers, reports, summaries). Discovering unsuspected links from the huge amount of literature and supporting scientific research (McInerney, 2002; Gomez-Perez et al., 2009)

This research paper considers the methodology of applying the methods of semantic analysis for revealing hidden connections between documents with the objective to classify and recognition the topical similarity.

2. Theoretical Justification

Under the notion of texts mining in natural language we understand the application of methods of texts computer analysis and presentation in order to achieve the quality, which corresponds to the "manual" processing for further usage in various tasks and applications. One of the actual tasks of automatic texts mining is their clustering (definition of groups of the similar documents). More and more often statistical topical methods are being applied (Vorontsov and Potapenko, 2013).

The topics are presented as discrete distributions on a number of words, and the documents – as discrete distribution on a number of topics (Vorontsov and Potapenko, 2013). Topical methods perform a "non-precise" clustering of words and documents, which means that a word or a document can be referred to a few topics with different probabilities simultaneously. The synonyms with higher probability will appear in the same topics since they are frequently used in the same documents. At the same time, the homonyms (words different in meaning, but similar in writing) will be placed in different topics because they are used in different contexts (Dempster et al., 1977).

Vector Space Model

Topical methods, as a rule, apply the method of a "bad of words", where each document is considered as a set of words not connected to each other. Before the topics are defined, the text is processed – its morphologic analysis is conducted with the objective to define the initial form of words and their meanings in the speech context. The method of processing words in a machine-readable natural language, as a rule, is based on the vector-space method of data description (Vector Space Model) (Nokel and Lukashevich, 2015), suggested by G. Solton in 1975. Within the framework of the method each word in a document has its particular weight. Thus, each document is presented as a vector and its dimension is equal to the total number of words in the document.

Similarity of a document and a topic is evaluated as a scalar product of a few information vectors. The weight of separate words (terms) can be calculated both applying the absolute frequency of a word appearing in the text and the relative (normalized) frequency:

$$tf(w,t) = \frac{k(w,t)}{df}, \tag{1}$$

where $k(w, L_t)$ is the number of *w*-word occurrences in the text *t*; *df* – total number of words in the text *t*;

The weight of a word, calculated by the formula (1), in documents is usually put as TF (Term Frequency).

However, this approach does not take into consideration the frequency, with which the word is used in the whole massive of documents – i.e. the so-called discrimination strength of the word. That is why, in case when the statistics for word usage in the whole document is available, it is more efficient to use the other method (formula 2):

$$TF \times IDF = tf(w,t) \cdot \log_2 \cdot D/df \tag{2}$$

where *D* – total number of documents in the collection.

$TF \times IDF$ method of weighting words shows not the frequency of words appearing in the document, but the measure, inverse to the number of documents in the massive containing this particular word (inverse document frequency).

The Vector Space Model of data presentation provides the systems, which are based on it, with the following functions:

- creation of professional systems and databases;
- increase of the level of specialists' competence by means of obtaining an effective possibility of directed search and filtration of text documents;
- automatic summarization of documents' texts.

Vector Space Model has:

- advantages: Index words can be selected automatically; word weighting to improve retrieval performance; partial matching of queries and documents when no document contains all search words; relevance ranking according to similarity; relevance feedback incorporated by modifying query vector;
- limitations the calculations used by simple vector models are about the frequency of words and word forms (e.g., stemmed) in texts; this means that they are measuring the "surface" usage of words as patterns of letters; they can't distinguish different meanings of the same word (polysemy); they can't detect equivalent meaning expressed with different words (synonymy); the vector model is not always suitable for solving the tasks of providing information security, because it refers to the methods of "hard" clustering (Nosenko et al., 2005), considering each document as a sparse vector in a space of words of big dimension.

Latent Semantic Analysis

In "soft" clustering each word and document refer to a few topics with particular probabilities simultaneously. Semantic description of a word or a document is a probability distribution on a number of topics. The process of finding these distributions is called "topical modelling".

One of the best methods of "soft" clustering is the Latent Semantic Analysis (LSA), which reflects documents and separate words in the so-called "semantic space", where all the further comparisons are conducted (Dumais et al., 1988; Deerwester et al., 1990).

In this process, the following assumptions are made:

- documents are a set of words, the order of which is ignored; it is only important how many times a word appears in the text;
- semantic meaning of a document is defined by the set of words, which, as a rule, go together;
- each word has a single meaning.

LSA is the method of processing information in natural language, analyzing interconnections between massifs of documents and words, appearing in them, as well as associates topics with documents (words).

The LSA method is based on principles of revealing latent connections of the studied phenomena and objects. In classification/clustering of documents this method is applied to extract context-dependent meanings of lexical units by means of statistical processing of very large text massifs. As the initial information in LSA the matrix "word-document" is used, which describes the set of data, used for system's training.

Elements of this matrix contain weights that consider frequencies of using every word in every document and participation of a word in all documents ($TF \times IDF$). The most

widely-used variant of LSA is based on using decomposition of a diagonal matrix by singular values (SVD – Singular Value Decomposition).

With the help of SVD any matrix can be decomposed on many orthogonal matrices, the linear combination of which is a rather precise approximation to the initial matrix.

Mathematical basis of the method is as follows:

Formally let *A* be a $m \times n$ words-document matrix of a documents collection. Each column of *A* corresponds to a document. The values of the matrix elements $A[i, j]$ represent the frequency identifications $tf(w,t)$ of the word occurrence w_i in the document t_j: $A[i, j] = tf(w,t)$. The dimensions of *A*, *m* and *n* correspond to the number of words and documents, respectively, in the collection.

In this case $B = A^T A$ is the document-document matrix. If the documents *i* and *j* have *b* words in common, then $B[i, j] = b$. On the other hand, $C = AA^T$ is the word-word matrix. If the words *i* and *j* occur together in *c* documents then $C[i, j] = c$. Clearly, both *B* and *C* are square and symmetric; *B* is a $m \times m$ matrix, whereas *C* is an $n \times n$ matrix. Now, we perform the Singular Value Decomposition on *A* using matrices *B* and *C* as described in the previous section:

$$A = U\Sigma V^T \tag{3}$$

where *U* is the matrix of the eigenvectors of *B*; *V* is the matrix of the eigenvectors of *C*; Σ is the diagonal matrix of singular values obtained as square roots of the eigenvalues of *B*.

In LSA we ignore these small singular values and replace them by 0. Let us say that we only keep *k* singular values in *Σ*. Then *Σ* will be all zeros except the first *k* entries along its diagonal. We can reduce the matrix *Σ* into Σ_k which is a $k \times k$ matrix containing only the *k* singular values that we keep, and also reduce *U* and V^T, into U_K and $V_k{}^T$, to have *k* columns and rows, respectively. Of course, all these matrix parts that we throw out would have been zeroed anyway by the zeros in Σ. Matrix *A* is now approximated by:

$$X_{t \times d} \approx X_{K_{t \times d}} = U_{K_{t \times d}} \Sigma_{K_{t \times d}} \left(V_{K_{t \times d}}\right)^T \tag{4}$$

Intuitively, the *k* remaining ingredients of the eigenvectors in *U* and *V* could be interpreted as a k "hidden concepts" where the words and documents participate. The words and documents now have a new representation in words of these hidden concepts. The results of the LSA method are following:

Document comparison: $Z_k = \Sigma_{K_{t \times d}} \left(V_{K_{t \times d}}\right)^T$ represents docs (cols) in semantic space (scaling with singular values). Documents d_i and d_j can be compared using cosine distance on *i* and *j* columns of Z_k.

Word comparison $Y_k = U_{K_{t \times d}} \Sigma_{K_{t \times d}}$ represents words (cols) in semantic space. Words t_i and t_j can be compared as cosine distance on *i* and *j* columns of Y_k.

Topic analysis. Left singular vectors $U_{K_{t \times d}}$ map between *k* words and "semantic dimensions" (topics). Then column *k* of this vector "describes" topic by giving strength of

association with each word. Right singular vectors $V_{K_{t\times d}}$ map between topics and documents could in principle tell us what a document was "about". As with words, one document can be associated with many topics.

The method has the following advantages:

- it is the best for revealing latent dependencies in documents;
- it can be applied both with and without education (e.g. in solving the tasks of clustering); the values of matrix of closeness, based on the frequency characteristics of documents and lexical units;
- polysemy and homonymy are partially eliminated.

Disadvantages of the method are follows:

- significant decrease of the speed of calculations when the volume of input data increases;
- the applied probability method does not always correspond with the reality. It is suggested that words and documents have normal distribution, though usage of the Poisson distribution is more real.

To get rid of the above-mentioned disadvantages the probability LSA is conducted, based on the multinomial distribution – in particular, on the algorithm of Latent Dirichlet allocation (LDA) (Blei et al., 2003; Blei, 2012).

The LDA presupposes that each word in a document is created by a certain latent topic; at the same time distribution of words in each of them is used in a clear form, as well as the prior distribution of words in the document. Topics of all the words in the document are supposed to be independent. In LDA, as well as in LSA, a document can correspond to a few topics. However, LSA sets the algorithm of generation of both words and documents, that is why there appears an additional possibility to evaluate probabilities of documents outside text massive using the algorithm of variation Gibbs sampling.

Unlike LSA, in the LDA the number of parameters does not increase with the growth of number of documents in the studied massive. The applied extensions of the LDA algorithm eliminate some of its limitations and improve productivity for particular tasks. LSA is generating algorithm only for words, but not for documents. The LDA algorithm overcomes this limitation.

The main idea of LDA consists in the fact that the documents are presented as a mix of distributions of latent topics, where each topic is defined by a probability distribution on the set of words. LDA reflects hidden connections between the words by means of topics; it also allows to set probabilities for new documents, which were not included into the training set, applying the algorithm of Variational Bayesian method.

In fact, LDA is a three-stage Bayesian network, which generates a document from a mix of topics. At the first stage for each document d a random vector θ_d with the parameter α (usually α is taken as 50/T) is selected from the Dirichlet distribution. At the second stage a topic z_{di} is selected form the multinomial distribution with the parameter θ_d. Finally, in accordance with the selected topic zd_i a word w_{di} is chosen from the distribution Φ_{zdi}, which is the Dirichlet distribution with the parameter β (usually the parameter β is 0,1; its increase leads to more sparse topics).

3. Experimental Results and Finding

As an example, demonstrating the possibilities of applying algorithms of semantic analysis for a knowledge worker of the education sphere, it is suggested to consider the task

of semantic analysis and thematic classification of scientific papers with the objective to regulate the process of text mining, necessary for the research.

The algorithm for the experiment is the following (Rizun et al., 2016a, 2016b, 2017):

1. Conduction of the procedure of text pre-processing.
2. Latent Semantic Analysis of texts with the objective to reveal hidden semantic connections between the documents. Formation of semantic clusters of documents.
3. Definition of topics of papers with appointing topics to separate papers. Clustering of documents on the basis of topical identity.
4. Additionally – statistical analysis of documents' authorship applying the second Zipf's law.

As the specific task of a knowledge worker the following is suggested:

To form the literature review it is necessary to conduct analysis of the major areas of interest of a group of authors. Further these results can be applied to define the target topics (define the weight of papers of each topic in the total scientific heritage of the authors). At the same time, it is known that the authors' works have the topics that can overlap. This analysis is difficult to conduct applying only the papers' titles. Each worker has to spend an unreasonably large amount of time to read the papers.

As the text sample, it is suggested to choose 10 scientific papers of the same authors working on two close scientific topics – modelling of a human decision-making process and semantics texts analysis.

Step 1. Pre-processing – includes the following procedures: extraction of words for the word-document matrix; tokenization; transformation into lower case letters; removal of stop words; removing the words, which occurred only once (optionally); stemming/lemmatization.

Step 2. On the basis of the obtained "bag of words" it is possible to build a word-document matrix by the further processing of a vector model of documents applying LSA.

As a measure of closeness between the documents the cosine distance was used:

Cosine of the edge between the vectors:

$$dist_{t_i} = \cos\alpha = \frac{x \cdot y}{\|x\| \cdot \|y\|}, \tag{5}$$

where $x \cdot y$ is a scalar product of the vectors, $\|x\|$ and $\|y\|$ – quota of the vectors, which are calculated by the formulas:

$$\|x\| = \sqrt{\sum_{i=1i}^{n} x_i^2}, \ \|y\| = \sqrt{\sum_{i=1i}^{n} y_i^2} \tag{6}$$

On the basis of the matrix of cosine distances between documents the clustering of analyzed documents was conducted. Results of the conducted research with changed clustering parameters are shown in Table 2.

Table 2.

LSA-based Clustering Results

Documents	2 clusters	3 clusters	4 clusters	Paper titles
0	1	0	1	Instruments of Operator's Active State Identification
1	1	0	1	Simulation Model of the Decision-Making Support for Human-Machine Systems Operators
2	1	0	2	Methodology of Intellectual Express-Evaluation of the Decision-Making Effectiveness
3	1	1	2	Models of Human Decision-Making Processes
4	1	0	1	Innovative Methods and Models of Individuals' Knowledge Identification
5	0	2	3	The Method of a Two-Level Text-Meaning Similarity Approximation of the Customers' Opinions
6	0	2	4	Text-mining Similarity Approximation Operators for Opinion Mining in BI tools
7	0	2	3	Development and Research of the Text Messages Semantic Clustering Methodology
8	0	2	3	Algorithm of the Hidden Semantic of the Films Reviews Analysis
9	0	2	3	Algorithm of Polish language Film Reviews Preprocessing

This table allows drawing conclusions about the fact that growth of the number of clusters allows to increase the precision of subjects' recognition – from the more general to the sub-topics:

- when transferring from two topics to three, the subject connected with modelling the processes of decision-making was divided into sub-topics. As a result, one document was put into a separate cluster (that paper really has a bigger share of mathematic and technical interpretations and the following classification of different models of human decision-making in various conditions);
- when transferring to four topics the hierarchy allowed to define also the sub-topics for a group of papers, dedicated to texts semantics analysis.

Graphic interpretation of the obtained results is presented in Figure 1.

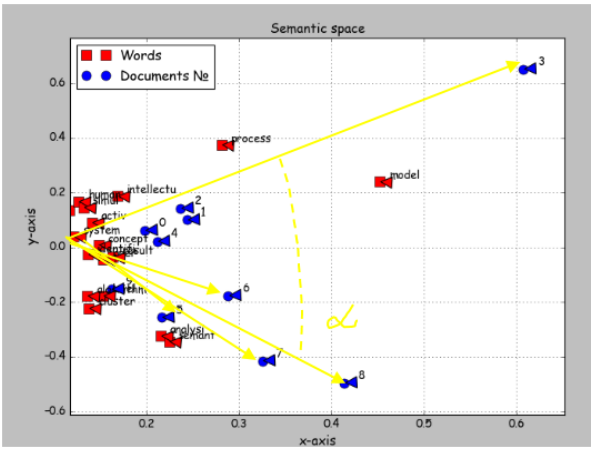

Figure 1. Graphic Interpretation of LSA-based Clustering Results

Step 3. The group of documents that were obtained as a result of semantic clustering and are characterized by topical closeness, require formulation of these topics as well as verification of precision of the conducted procedure taking into consideration the above-mentioned disadvantages of the method.

The experiment on modelling the topics was performed by means of the latent placement of Dirichlet (LDA). The most common method of evaluating the quality of probabilistic topic models is calculating the perplexity index on the test data set $D_{tes}t$ of M documents (Bahl et al., 1977).

In information theory, perplexity is a measurement of how well a probabilistic distribution or probabilistic model predicts a sample. It may be used to compare probabilistic models. A low perplexity indicates that the probabilistic distribution is good at predicting the sample:

$$Perplexity(D_{test} = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (7)$$

Documents classification was conducted on the basis of the index of maximum probability of the document belonging to the particular topic. One of the major factors influencing the quality of topics definition is the ratio: number of words in the topic / total number of topics. When this index decreases (→1) the perplexity index decreases as well. Results of the conducted research with changing the number of words in a topic as well as the number of topics, are given in Tables 3, 4 and 5. Key words of each topic are presented in the form obtained after the pre-processing – i.e. stemming.

Table 3.

LDA-based Clustering Results (4 topics/clusters)

Topics	**Key words of topics**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Topic 3	analysi semant level cluster						+		+	+	+
Topic 2	oper system model proces			+				+			
Topic 1	model intellectu process simul				+	+					
Topic 0	activ identif state result	+	+								
Probability		0,59	0,71	0,66	0,71	0,57	0,56	0,59	0,83	0,79	0,77
Perplexity=7592		Optimal number of words in a topic – 4									

Table 4.

LDA-based Clustering Results (3 topics/clusters)

Topics	**Key words of topics**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Topic 2	model activ identif	+				+					
Topic 1	semant cluster analysi						+	+	+	+	+
Topic 0	algorithm individu model		+	+	+						
Probability		0,71	0,61	0,54	0,66	0,60	0,81	0,77	0,79	0,84	0,78
Perplexity=2064		Optimal number of words in a topic – 3									

Table 5.

LDA-based Clustering Results (2 topics/clusters)

Topics	**Key words of topics**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Topic 0	semant analysi						+	+	+	+	+
Topic 1	model activ	+	+	+	+	+					
Probability		0,82	0,84	0,77	0,79	0,67	0,71	0,89	0,83	0,75	0,77
Perplexity=574		Optimal number of words in a topic – 2									

Comparison of the results obtained at stages 2 and 3 of the documents semantic analysis are presented in Table 6.

Table 6.

Comparison of the LDA- and LSA- based Clustering Results

Documents	2 topics/clusters			3 topics/clusters			4 topics/clusters		
	LSA	LDA	Topics	LSA	LDA	Topics	LSA	LDA	4 Topics
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
0	1	1	Model of Activity	0	0	Individuals Knowledge Identification	1	1	Active State Identification Results
1	1	1	Model of Activity	0	2	Model of Activity Identification	1	1	Active State Identification Results
2	1	1	Model of Activity	0	2	Model of Activity Identification	2	2	System of Operator's Processes Models
3	1	1	Model of Activity	1	2	Model of Activity Identification	2	0	Model of Intellectual Process Simulation
4	1	1	Model of Activity	0	0	Individuals Knowledge Identification	1	0	Model of Intellectual Process Simulation
5	0	0	Semantic Analysis	2	1	Semantic Clustering Analysis	3	3	Analysis of Semantic Clusters Level
6	0	0	Semantic Analysis	2	1	Semantic Clustering Analysis	4	2	System of Operator's Processes Models
7	0	0	Semantic Analysis	2	1	Semantic Clustering Analysis	3	3	Analysis of Semantic Clusters Level
8	0	0	Semantic Analysis	2	1	Semantic Clustering Analysis	3	3	Analysis of Semantic Clusters Level
9	0	0	Model of Activity	2	1	Individuals Knowledge Identification	3	3	Analysis of Semantic Clusters Level
% of results discrepancy = 0%				% of results discrepancy = 20%			% of results discrepancy = 30%		

These results prove that the optimal variant of dividing publications on topics is with the 2 groups. However, when it is necessary to clarify and divide the overlapping topics, we can consider 3 clusters of analyzed documents.

Besides, in addition it is possible to conduct the test of the authorship of publications.

In a short publication (mignews.com.ua, 2009) it was stated that in the scientific edition "New Journal of Physics" a group of Swedish physics from the Umea university under the direction of Sebastian Bernhardsson has described a new method that allows to define the author of a text on the basis of statistical data. The researchers have checked how in the texts of three authors – Thomas Hardy, Henry Melville and David Lawrence, – the so-called law of Zipf is realized. The researchers have detected that the frequency of new words appearance as the volume of a text grows is changing differently for each author, and this

pattern does not depend on a particular text – only on the author. This information was published in 2009, and more than 20 years ago John Charles (Baker J. C. Baker, 1998) introduced a unit for measuring the capability of an author to use new words (here the notion "new" means a word not used in the text before). Baker proved that the unit is the author's individual characteristics.

Figure 2 shows an example of three publications (of one size) comparison. Moreover, the first two papers (Text 1and Text 2) are written by the same group of authors, and the third (Text 3) – by another author.

Table 7.

Parameters of the Zipf's Law Distribution

Document	Factor *a*	Factor *b*	Mistake of approximation	Total number of words	New words	% of new words	Average distances between texts	
1	5.641	71.172	0.5795%	1974	670	34	1st and 2nd	1st and 3rd
2	6.332	58.466	0.5417%	2028	682	34	0.993	1.389
3	4.371	70.154	0.4592%	1951	806	41		

The curves of the power distribution (parameters of the function $y = a + b/x$ are given in Table 7), built as the approximation of the observed regularities, clearly demonstrate the differences between authors' "fingerprints":

- the distance between 1st and 2nd text is bigger than between 1st and 3rd;
- % of new words in the texts 1 an 2 approximately the same.

This methodology allows to make sure that Zipf's "quantity-frequency" law really can be applied for any text written in any language. Moreover, in this example, Zipf's law was built after the procedure of text pre-processing was conducted. That enables a more precise analysis of the vocabulary and the specificity of authors using new words.

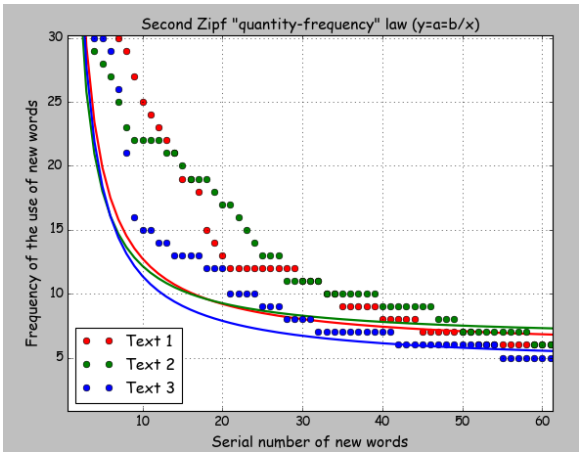

Figure 2. Zipf's Law Distributions for Texting the Authorships

Conclusions

Thus, different aspects of applying the methodologies of text analysis for realization of knowledge worker's activity were analyzed. We can draw conclusions on the urgency and importance of using the methodologies of processing the non-structured textual information for increasing the efficiency of knowledge workers, as well as their awareness in different spheres of economics.

The paper analyses the existing algorithms of texts semantic analysis as the sphere of documents topical closeness recognition. It was proved that these algorithms allow knowledge workers to solve various routine as well as creative tasks on a high intellectual level.

The example given in the paper describes the complex methodology of semantic analysis for a combination of LSA and LDA methods and application of Zipf's law for solving one of the typical knowledge worker's tasks. Using this methodology allows to reduce the time on documents processing to 40%, as well as guarantees the matching of results of topics definition with the results of human work – to 70-75%.

Literature

Bahl L., Baker, J., Jelinek E. & Mercer, R. (1977) Perplexity – a Measure of the Difficulty of Speech Recognition Tasks. In Program, 94th Meeting of the Acoustical Society of America, volume 62, page S63.

Baker, J.C. (1998) A Test of Authorship Based on the Rate at which New Words Enter an Author's Text. Journal Article published 1 Jan 1988 in Literary and Linguistic Computing, volume 3, issue 1, pp. 36-39.

Blei, D. M. (2012) Introduction to Probabilistic Topic Models. Comm. ACM 55 (4), April, 2012: pp. 77-84

Blei, D. M., Ng, A. & Jordan, M. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: pp. 993–1022.

Bose, R. P. J. C. & and van der Aalst, W. M. P. (2009) Context Aware Trace Clustering: Towards Improving Process Mining Results. In SIAM International Conference on Data Mining, pages 401–412.

Cantú, F.J. & Ceballos, H.G. (2010) A Multiagent Knowledge and Information Network Approach for Managing Research Assets. Expert Systems with Applications, 37(7), 5272-5284. doi:10.1016/j.eswa.2010.01.012

Cheng, H., Lu, Y. & Sheu, C. (2009) An Ontology-Based Business Intelligence Application in a Financial Knowledge Management System. Expert Systems with Applications, 36, 3614–3622. Doi:10.1016/j.eswa.2008.02.047

Deerwester, S., Dumais, S. T. & Harshman, R. (1990) Indexing by Latent Semantic Analysis. http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B., Vol. 39. No 1, pp. 1-38

Dumais, S. T., Furnas, G. W., Landauer, T. K. & Deerwester, S. (1988) Using Latent Semantic Analysis to Improve Information Retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285

Ghattas, J., Peleg, M., Soffer, P., & Denekamp, Y. (2010) Learning the Context of a Clinical Process. In Business Process Management Workshops, pages 545–556. Springer.

Gomez-Perez, J. M., Grobelnik, M., Ruiz, C., Tilly, M., & Warren, P. (2009) Using Task Context to Achieve Effective Information Delivery. In Proceedings of the 1st Workshop on Context, Information and Ontologies, pages 1–6. ACM.

Hwang, H.G., Chang, I.C., Chen, F.J. & Wu, S.Y. (2008) Investigation of the Application of KMS for Diseases Classifications: A Study in a Taiwanese Hospital. Expert Systems with Applications, 34(1), 725-733. doi:10.1016/j.eswa.2006.10.018

Li, X., Zhu, Z. & Pan, X. (2010, a) Knowledge Cultivating for Intelligent Decision Making in Small & Middle Businesses. Procedia Computer Science, 1(1), 2479-2488. doi:10.1016/j.procs.2010.04.280

Li, Y., Kramer, M.R., Beulens, A.J.M. &Van Der Vorst, J.G.A.J. (2010, b) A Framework for Early Warning and Proactive Control Systems in Food Supply Chain Networks. Computers in Industry, 61, 852–862. Doi:101.016/j.compind.2010.07.010

Liao, S. (2003) Knowledge Management Technologies and Applications-Literature Review From 1995 to 2002. Expert Systems with Applications, 25, 155-164. doi:10.1016/S0957-4174(03)00043-5

Liao, S.H., Chen, C.M. and Wu, C.H. (2008) Mining Customer Knowledge for Product Line and Brand Extension in Retailing. Expert Systems with Applications, 34(3), 1763-1776. doi:10.1016/j.eswa.2007.01.036

Liu, D.R. & Lai, C.H. (2011) Mining Group-Based Knowledge Flows for Sharing Task Knowledge. Decision Support Systems,50(2), 370-386. doi:10.1016/j.dss.2010.09.004

McInerney, C. (2002) Knowledge Management and the Dynamic Nature of Knowledge. Journal of the American Society for Information Science and Technology, 53(12), 1009-1018. doi:10.1002/asi.10109

MIGnews.com.ua (2009) Authorship of writers can be learned by a special formula, http://mignews.com.ua/science/nauka/2531956.html

Nokel, M. A. & Lukashevich, N.V. (2015) Thematic Models: Adding Bigrams and Accounting Similarities Between Unigrams and Bigrams. Computational methods and programming. Vol. 16, pp. 215-217

Nosenko, S. V, Korolev, I. D. & Poddubny, M.I. (2005) The Method of Automatic Classification Formalized Documents in the System of Electronic Document Management, MKK G06F17 / 30

Rizun, M. (2017a) Maturity Models as the Element of Knowledge Management Development. Materials of the Conference „Current Issues Raised by Young Researchers, X". CreativeTime, Krakow. ISBN 9788363058-71-5. Pp. 293 – 298.

Rizun, M. (2017b) Software Solutions for Knowledge Workers. Proceedings of the 2nd International Conference on Information Technologies in Management, Publisher: Rocznik Naukowy Wydziału Zarządzania WSM, http://www.wsmciechanow.edu.pl/rocznik-naukowy/ (in print).

Rizun, N. & Taranenko, Y. (2017) Development of the Algorithm of Polish Language Film Reviews Preprocessing. Proceeding of the 2nd International Conference on Information Technologies in Management, Publisher: Rocznik Naukowy Wydziału Zarządzania WSM, http://www.wsmciechanow.edu.pl/rocznik-naukowy/ (in print).

Rizun, N., Kapłanski, P. & Taranenko, Y. (2016a) Development and Research of the Text Messages Semantic Clustering Methodology. 2016, Third European Network Intelligence Conference, Publisher: ENIC, # 33, pp.180-187.

Rizun, N., Kapłanski, P. & Taranenko, Y. (2016b) Method of a Two-Level Text-Meaning Similarity Approximation of the Customers' Opinions. Economic Studies – Scientific Papers. University of Economics in Katowice, Nr. 296/2016, pp.64-85.

Silwattananusarn, T. & Tuamsuk, K. (2012) Data Mining and its Applications for Knowledge Management: A Literature Review from 2007 to 2012, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012, pp.13-24

Stajner, T. & Mladenic, D. (2010) Modeling Knowledge Worker Activity. MLR: Workshop on Applications of Pattern Analysis. Workshop and Conference Proceedings 11 (2010), pp. 127–133.

Vorontsov, K.V. & Potapenko, A.A. (2013) Modifications of the EM-algorithm for probabilistic Thematic modeling // Machine learning and data analysis. Vol. 1. No. 6, pp. 657-686

Wang, F. & Fan, H. (2008) Investigation on Technology Systems for Knowledge Management.IEEE, 1-4. doi:10.1109/WiCom.2008.2716

Wu, W., Lee, Y.T., Tseng, M.L. & Chiang, Y.H. (2010) Data Mining for Exploring Hidden Patterns Between KM and its Performance. Knowledge-Based Systems, 23, 397-401. doi:10.1016/j.knosys.2010.01.014