Approaches to assessing the semantic similarity of texts in a multilingual space

A.Kh. Khakimova[1], M.M. Charnine[2], A.A. Klokov[2], E.G. Sokolov[2]
aida_khatif@mail.ru | mc@keywen.com | aaklokov@yandex.ru | evgeny.sokolov@phystech.edu
[1]ANO «Scientific and Research Center for Information in Physics and Technique», Nizhny Novgorod, Russia;
[2]FRC CSC of the Russian Academy of Sciences, Moscow, Russia

This paper is devoted to the development of a methodology for evaluating the semantic similarity of any texts in different languages is developed. The study is based on the hypothesis that the proximity of vector representations of terms in semantic space can be interpreted as a semantic similarity in the cross-lingual environment. Each text will be associated with a vector in a single multilingual semantic vector space. The measure of the semantic similarity of texts will be determined by the measure of the proximity of the corresponding vectors. We propose a quantitative indicator called Index of Semantic Textual Similarity (ISTS) that measures the degree of semantic similarity of multilingual texts on the basis of identified cross-lingual semantic implicit links. The setting of parameters is based on the correlation with the presence of a formal reference between documents. The measure of semantic similarity expresses the existence of two common terms, phrases or word combinations. Optimal parameters of the algorithm for identifying implicit links are selected on the thematic collection by maximizing the correlation of explicit and implicit connections. The developed algorithm can facilitate the search for close documents in the analysis of multilingual patent documentation.

***Keywords**: cross-lingual semantic similarity, semantic textual similarity measure, semantic implicit links, collection of documents, measure of similarity of texts, method of relevant phrases, vector representations for words.*

1. Introduction

As cross-language information retrieval gets more attention, tools to measure cross-language semantic similarity between documents become necessary. An accurate assessment of the actual similarity between documents is fundamental for many automatic text analysis applications, such as thesaurus generation [1], machine translation [2], information search [3], automatic generalization [4].

Text mining and knowledge management technologies play a key role in many areas, including critical infrastructures. Information search, document classification, business analytics, forecasting technologies, etc. are currently the most important activities.

Patent search, including monitoring competitors, checking the novelty of an invention, or searching for technical solutions in other fields of application, requires a lot of effort.

Comparing documents in different languages is challenging for natural language processing applications, and especially in machine translation applications.

Cross-language matching of documents is carried out in a patent search to protect an invention in more than one country or region. A separate patent must be filed with several patent offices in different languages. Before applying for a patent, applicants conduct a preliminary search for patents or documents revealing intellectual property similar to the filed invention. In such a process, a set of patents is requested in one language, using the source document in another language as a request.

To compare the received documents, it is necessary to use cross-language similarity assessment functions. This task can be formulated as discarding text pairs that are not semantically equivalent [5]. The task is complicated by the fact that in the case of filing an invention in different countries, different standards may be used, which may lead to a discrepancy between versions of the document in different languages. In this case, the task of identifying semantic equivalents is complicated [6].

Natural language processing methods for text analysis and data mining are used in the analysis of many types of technical documentation. Functional analysis methods are based on extracting interactions between the entities described in the document.

Linguistic analysis tools permit to identify key elements of a document by combining morphological, syntactic, and semantic analysis. Application of methods of linguistic analysis to patent documents allows for accelerated analysis and comparison of patents.

The purpose of the analysis of technical documentation is to discover possible ambiguities or incompleteness on the one hand, and understanding the requirements in the direction of possible formalization on the other.

The main problem here is that keyword searches do not take into account synonyms or more abstract terms associated with given query words. This means that if a synonym is used for an important term in a patent application, for example, a wire instead of a cable, a keyword search may not reveal this relationship if an alternative term was not explicitly included in the search query. This is relevant since patent texts often use abstract and general terms to describe the invention in order to maximize protection [7].

If we consider the Internet as a multilingual database, a typical problem when searching for information is the search for relevant documents in the collection of documents by some key terms, or by the example of the corresponding document. Assessing the semantic similarity between words (phrases) is critical to assessing whether a document meets user needs. Many information retrieval systems, such as online library catalog systems, web search engines, deal with multilingual documents and must have tools to measure cross-language semantic similarity.

In recent decades, many studies have been carried out aimed at improving the effectiveness of measures of semantic similarity of words. However, studies of semantic similarity mainly focus on English. This is partly

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR-WS.org/Vol-2763/CPT2020_paper_s3-9.pdf

due to the limited availability of similarity criteria for words in languages other than English. Since the development of multilingual methods is necessary, there is an urgent need to find a reliable basis for assessing multilingual and interlingual semantic similarity.

Despite the fact that in many areas a multilingual measurement of semantic semantic similarity is required, most algorithms measure semantic similarity between words of the same language. Cross-language similarity was first described in 2009 [8] for Anglo-Spanish cross-language data sets. Over the past few years, multilingual word embeddings, which are lexical elements from several languages in a single semantic space, have attracted considerable attention of researchers [9-11].

Interlanguage applications are based on data mining methods, such as text clustering, which includes extracting words or phrases from documents as functions, representing documents as feature vectors, and then grouping documents into clusters based on similarity of feature vectors. In a multilingual document collection, recoverable functions will refer to multilingual words. Therefore, it is important to measure the similarity between the words of not only one language, but also of different languages.

According to the concept of the information data space [12], the information space should model a rich set of relationships between data repositories. To model the relationship between data warehouses in data spaces, you need a component that can measure the semantic similarity between interlanguage pairs. Sources in a data space can be relational databases, XML repositories, text databases, web services, etc.

The problem of plagiarism in a monolingual context is well developed [13]. Free machine translation tools help spread cross-language plagiarism (plagiarism by translation). In this relatively new field of research, the definition of semantic text similarity in language pairs has been carried out. The authors investigated various existing approaches to detect plagiarism on different language pairs and found that if the method is effective for a particular language pair, it will be equally effective for another language pair with a sufficient number of available lexical resources, i.e. the method can be optimized for a particular case and is effectively applied on another case [14].

2. Methodology for calculating the assessment of semantic similarity

The technique includes the following steps:

1) pre-processing of texts by replacing their terms with synset codes;
2) construction of quotation vectors by identifying common rare phrases (long quotes) in various documents using the relevant phrases method;
3) thematic analysis of processed texts and building a set of available topics and corresponding thematic document vectors using the LDA method with the possibility of further clustering documents on topics / ideas into “baskets”/clusters;
4) the construction for each document of an extended vector describing the presence of long citations, the statistics of the synsets included in it and their thematic composition, i.e. the document vector is the concatenation of the citation vector, thematic vector and synset statistics vector;
5) calculation of the similarity index between articles/documents (Semantic Text Similarity Index, ISTS) by the cosine measure of the corresponding article vectors;
6) calculation of the correlation between the formal connectedness of articles and their similarity index, taking into account the minimum and maximum thresholds of the ISTS;
7) the choice of values of various calculation parameters (ISTS thresholds) based on the maximum correlation.

The calculation method is selected according to the maximum correlation of ISTS with formal links.

In the basis of the algorithm for vector transformation of terms used recurrent neural networks (RNN - Recurrent neural network) - Fig. 1.

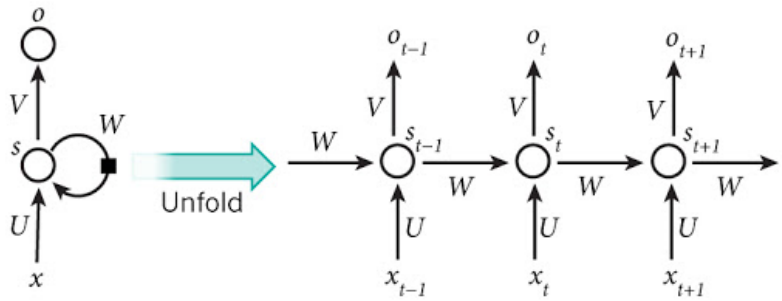

Fig. 1. Graphs of the number of points for calculating the correlations of the current and future years for the indicators IFTm (upper) and IFT, depending on the number of articles with the word in the last 3 years

RNN is used for tasks where there is a sequence of words and phrases. Formally, at each step (after each new processed word), RNN considers for each word in the corpus the probability of which word will be next. In this work, LSTM neurons, which are a special case of RNN, were used. Moreover, bi-directional recurrent biLSTM network (Bidirectional recurrent neural networks) was used. biLSTM is a combination of two LSTM networks in which at the same time one network builds a language model from the beginning of the sentence, and the second from the end.

We used the simplest sequential model, consisting of two layers. For the software implementation of the proposed architecture in Python, the jupyter notebook development environment was used. A linear layer was

attached to the biLSTM layer to solve the classification problem (Fig. 2).

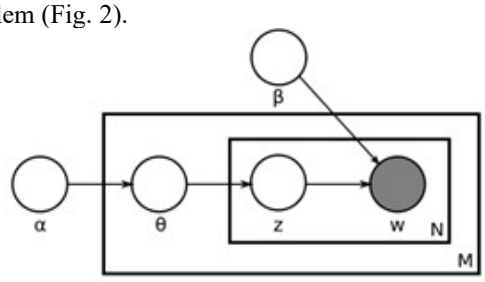

Fig. 2. Scheme of the LDA model

At the input of the neural network, vector representations of words (embending) were applied. Word2Vec was used to convert each word from the title of the article to a number vector. In the experiments, 300 dimension vectors were used (Word2Vec from the gensim library allows changing the embedding dimension).

In our experiments, we consider the DBLP citation network, a collection of articles on artificial intelligence compiled by aminer.org. In this study, we intentionally relied only on the title of the publication and its links. During the experiments, various models of the neural network were tested. Experiments were conducted with a change in the number of neurons in the biLSTM layer (4, 8, 16, 32, 64, 128) and the number of neurons in the linear layer (from 0 to 10). The best model was able to give an accuracy of 0.6131 according to the ROC AUC metric. The time for calculating the forecast and evaluating its accuracy was about 1 hour.

To combine articles with similar topics into clusters, we used generally accepted approaches to machine word processing (NLP), clustering articles using the Latent Dirichlet Allocation (LDA) method, and visualizing the results obtained with Python libraries. After extracting the data, preprocessing it, extracting tokens, stamping and deleting stop words, we used the Latent Dirichlet Allocation (LDA) algorithm - Fig. 2.

LDA is a hierarchical Bayesian model that consists of two levels: at the first level, a mixture whose components correspond to "themes"; at the second level, a multinomial variable with an a priori Dirichlet distribution that defines the "distribution of topics" in the document.

The principle of the model:

1) select the document length N
2) vector is selected θ ~ (α) - the vector of the "degree of expression" of each topic in this document;
3) for each of N words w:
   - choose a theme z_n by distribution $Mult(\theta)$;
   - choose a word $w_n \sim p(w_n|z_n, \beta)$ with probabilities given in β.

For simplicity, we fix the number of topics k and assume that β is just a set of parameters $\beta_{i,j} = p(w^j=1|z^i=1)$, which need to be evaluated, and we won't worry about the distribution on N. The joint distribution then looks like this:

$$p(\theta, , , N \mid \alpha, \beta) = p(N \mid \xi)p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta)p(w_n \mid z_n, \beta).$$

Unlike conventional clustering with an a priori Dirichlet distribution, we do not select a cluster here once, and then we look for words from this cluster, but for each word we first select a topic from the distribution θ, and only then we relate this word to this topic.

At the output after training the LDA model, themed vectors θ are obtained, showing how topics are distributed in each document, and distributions β, which show which words are more likely in certain topics. In our case, we got 8 pronounced clusters corresponding to the following directions:

1) computing systems and algorithms in them;
2) bioinformatics and data processing methods in it;
3) signal processing;
4) optimization methods and algorithms based on them;
5) problems related to theoretical informatics and computational complexity;
6) neural and computing networks;
7) issues regarding natural language processing (NLP) and programming languages;
8) robotics, and self-learning systems (Reinforcement Learning).

After the previous step, n-dimensional thematic vectors of articles are obtained. To compress the results into a two-dimensional vector space, the t-SNE machine learning algorithm was used. To visualize the clusters, we used an interface written in JavaScrIFT (Fig. 3).

The previous approach was based on a comparison of vectors at the megalemma level in a cosine measure, which determined the semantic similarity of the texts. As a development of this approach, based on the assumption that while maintaining the semantic similarity of phrases, ideas in them can be expressed in different words, we use the Impact Factor of the Term (IFT) to assess the similarity of documents.

To compare articles expressing new ideas, we use the hypothesis that new ideas are often expressed in terms of a high impact factor IFT. IFT is determined by the average number of links to articles with this term, the higher the IFT, the higher the citation trend and the number of formal links. If a couple of articles have a general term with a high IFT, the probability of a formal link between them will be high.

Using multilingual synsets built for high IFT terms (IFT terms), you can evaluate the similarity of articles in any language. If there is a semantic similarity, estimated by a cosine measure, it can be assumed that articles with this term will be quoted with some probability.

If previously the similarity of the vectors of megalemma determined the similarity of texts, now we use extended vectors based on common rare phrases, megalemmas and multilingual IFT synsets, as well as the results of thematic analysis. The similarity of extended vectors more accurately reflects the similarity of texts, since it takes into account not only semantic, but also thematic similarity.

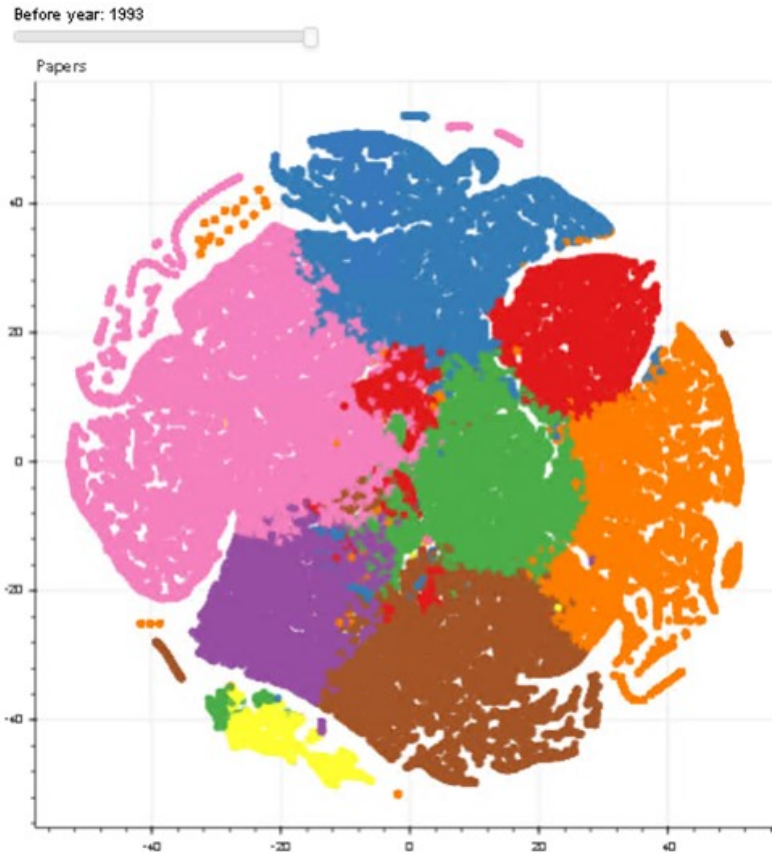

Fig. 3. Cluster states in 1993. 1) computing systems and algorithms in them (pink); 2) bioinformatics and data processing methods in it (purple); 3) signal processing (brown); 4) optimization methods and algorithms based on them (green); 5) problems related to theoretical informatics and computational complexity (orange); 6) neural and computing networks (red); 7) issues regarding natural language processing (NLP) and programming languages (blue); 8) robotics, and self-learning systems (Reinforcement Learning) (dark orange); yellow - a "garbage" cluster with articles in German

Our study is based on a model for representing ideas in the form of many terms and similar phrases in a multilingual semantic field, on the hypothesis that the proximity of vector representations of terms in a multilingual vector semantic space can be interpreted as semantic similarity in an interlanguage environment. We propose a method of formalizing ideas by using terms with high IFT and megalemma, which allows you to recognize an idea expressed in different words. References, both formal (bibliographic) and contextual (implicit, expressed by matching IFT terms), are an expression of the connection between ideas.

High IFT terms are significant terms (or ideologically significant). If the texts on the IFT synsets have the same vector, then this means the presence of common ideas in these texts and a significant similarity related to citation. The similarity in vectors of megalemmas also correlates with formal links (as our previous experiments showed), but to a much lesser extent. It is shown that megalemma has a very low impact factor.

It should be noted that the similarity in vectors of megalemmas is more applicable to texts with common vocabulary, in this case, the degree of coincidence of their thematic composition as a set of popular words is calculated. The approach to calculating the similarity of IFT / megalemma vectors is focused on comparing the similarity of scientific texts with specific terminology, despite the fact that ideas can have different lexical expressions. Therefore, in the second case, it becomes possible to more accurately assess the similarity from the point of view of ideological similarity, since terms with a high IFT are significant terms denoting ideas.

Three types of semantic similarity can be considered (based on implicit references): 1) similarity of the thematic composition of popular / common words (word frequency from 10 thousand or more); 2) the presence of common significant IFT terms denoting specific ideas (frequency 5-1000); 3) the presence of common rare phrases (long quotation) (frequency 2-100). These types differ in the frequency of matching terms / phrases. The highest frequency is typical for popular terms and megalemmas, the lowest is for common rare phrases. The proposed similarity assessment algorithm takes into account all these types of similarities, giving appropriate weights. Thus, when identifying similarities and implicit references, the entire frequency range of terms and phrases is used.

So, we build extended vectors from megalemmas and multilingual IFT synsets, and these can be weighted vectors whose elements have weights. The larger the impact factor, the higher the likelihood of a formal link and the higher the weight of the vector element. The cosine measure allows you to work with weighted vectors, in which elements take large real values. Since our task is to search for semantic similarity of articles correlating with the presence of formal links, then increasing the weights

of IFT synsets in extended vectors improves the quality of the proposed algorithm.

Therefore, the algorithm for calculating ISTS is based on assessing the similarity of vectors, expanded by adding multilingual IFT synsets and weights, according to a cosine measure, in order to determine the similarity of texts. This takes into account the presence of formal links between texts containing matching IFT terms. The method may contain options that are determined/selected by the optimization method according to the maximum correlation of ISTS with formal links.

The first version of the methodology for calculating the multilingual Index of Ideological Influence (III) as the number of similar subsequent / future articles / documents has been developed.

We consider similar subsequent articles to be articles that will cite this document, i.e. those articles are similar that are linked by formal links. Thus, the III is looking for trending articles containing trending IFT terms. We can calculate the second-level III, since one idea gives rise to another, then you can search for articles similar to the articles found in the first stage (indirect similarity / similarity). The mutual influence of articles is calculated using the PageRank algorithm [15], which increases the significance / influence of texts / articles the more they have more (implicit) links with other significant / influential texts.

IFT terms in scientific articles have an expiration date. The value of IFT is higher in the first years (3-4 years), and then it decreases (Fig. 4).

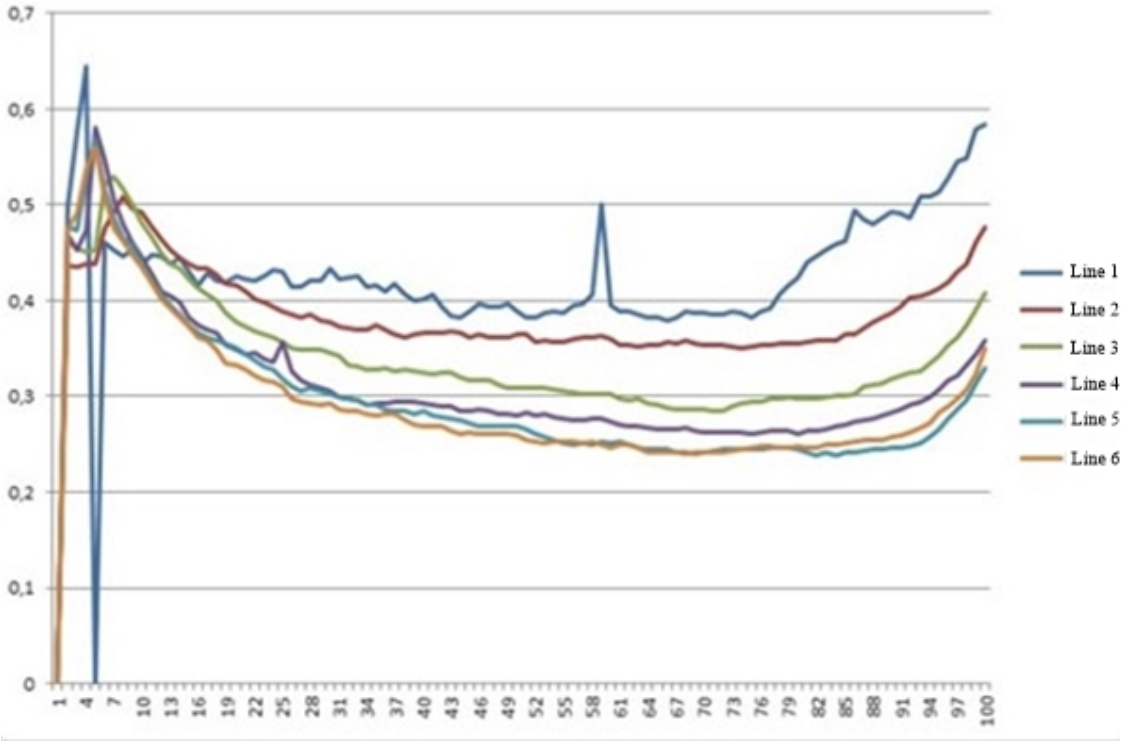

Fig. 4. Graphs of the average values of the IFT term, depending on the number of articles with these terms and the speed of the trend. 1 - 0 years, 2 - 1 year, 3 - 2 years, 4 - 3 years, 5 - 4 years, 6 - 5 years

Over time, some important terms are replaced by others. If in the vector for the term, in addition to the IFT, another year is introduced when the term was of high importance, then you can also obtain some information about the age of the article by the vector, which will allow you to find general ideas of a certain age when comparing the articles. This provides information on the dynamics of the development of ideas. For example, the term NEURAL NETWORKS has a long history, and in different years, various derivatives of this term were significant IFT terms, for example, FUZZY NEURAL or RECURRENT neural networks.

Over time, some important terms are replaced by others. If in the vector for the term, in addition to the IFT, another year is introduced when the term was of high importance, then you can also obtain some information about the age of the article by the vector, which will allow you to find general ideas of a certain age when comparing the articles. This provides information on the dynamics of the development of ideas. For example, the term NEURAL NETWORKS has a long history, and in different years, various derivatives of this term were significant IFT terms, for example, FUZZY NEURAL or RECURRENT neural networks.

So, the methodology for calculating the III contains the following steps:

1) search in the article for significant IFT terms;
2) compiling multilingual IFT synsets for these IFT terms;
3) on the basis of IFT-synsets, the definition of the forecast (regression analysis according to previous values of IFT and trend parameters);
4) refinement of the forecast using the PageRank algorithm [12], which increases the significance/influence of texts / ideas, the more they have (implicit) connections with other significant / influential texts.

In this case, implicit links between texts/articles are determined using the methodology for calculating the index of semantic text similarity (ISTS).

3. Results

As a result, we see the following pattern: the higher the forecast of the IFT, the higher the III of the document. The predictive value of the IFT is the same as the text, term, or idea. If there are several IFT terms in the text, then you can make a prediction according to the most significant/high IFT, or according to statistics that take into account the synergy of IFT terms when found together. An updated forecast of III/IFT is carried out using regression analysis using a number of indicators for the current year (IFT, IFTm, external links) and similar indicators of previous years.

4. Conclusion

The Multilingual Index of Ideological Influence (III) corresponds to the number of subsequent/future articles/documents citing the source document that are similar to the source document. We plan to consider a number of index modifications taking into account the cascade of citation (first and other levels) and the temporal dynamics of the development of ideas. It is planned to develop an algorithm for the updated forecast of III/IFT using a number of indicators of the current year (IFT, IFTm, external links) and similar indicators of previous years.

Acknowledgment

The reported study was funded by RFBR according to the research projects № 18-07-00909, 19-07-00857 and 20-04-60185.

References

[1] Jarmasz, M., Szpakowicz, S. (2003). Roget's Thesaurus and Semantic Similarity. Recent Adv. Nat. Lang. Process. III Sel. Pap. from RANLP 2003, vol. 111, 2004.

[2] Islam, A., Inkpen, D. (2012). Unsupervised Near-Synonym Choice using the Google Web 1T. ACM Trans. Knowl. Discov. Data, vol. V, no. June, pp. 1-19.

[3] Li, H., Xu, J. (2014). Semantic matching in search. Foundations and Trends in Information Retrieval, 7(5):343-469.

[4] Aliguliyev R. M. (2009). A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications. 36. 7764-7772. 10.1016/j.eswa.2008.11.022.

[5] Wäschle, K. (2015). Quantifying Cross-lingual Semantic Similarity for Natural Language Processing Applications. Heidelberg. – 139 p.

[6] Wäschle, K. and Riezler, S. (2012). Structural and topical dimensions in multi-task patent translation. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL).Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 818–828, Avignon, France, April 23 - 27, 2012

[7] Andersson, L., Hanbury, A. and Rauber, A. (2017). The Portability of Three Types of Text Mining Techniques into the Patent Text Genre, chapter 9, pages 241–280. Springer Berlin. Heidelberg, Berlin, Heidelberg. ISBN 978-3-662-53817-3.

[8] Eneko, A., Enrique, A., Keith, H., Jana, K., Marius, P., & Aitor, S. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 19-27). Boulder, Colorado: Association for Computational Linguistics

[9] Zou, W. Y., Socher, R., Cer, D.M. and Manning C.D. (2013). Bilingual word embeddings for phrase-based machine translation. In Proceedings of EMNLP (pp. 1393-1398).

[10] de Melo, G. (2015). Wiktionary-based word embeddings. Proceedings of MT Summit XV (pp. 346-359).

[11] Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C. and Smith, N.A. (2016). Massively multilingual word embeddings. arXiv preprint arXiv:1602.01925.

[12] Michael, J. F., Alon, Y. H., & David, M. (2005). From databases to data spaces: A new abstraction for information management. SIGMOD Record, 34(4), 27-33

[13] Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P. and Stein, B. (2014). Overview of the 6th International Competition on Plagiarism Detection. In PAN at CLEF 2014. Sheffield, UK (pp. 845-876).

[14] Ferrero, J., Besacier, L., Schwab, D. & Agnes, F. (2017). Using Word Embedding for Cross-Language Plagiarism Detection. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, (EACL 2017). Association for Computational Linguistics, Valencia, Spain, volume 2 (pp. 415-421).

[15] Page, L., Brin, S., Motwani, R., Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. In: Technical Report. Stanford University, Stanford, 1998. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

About the autors

Khakimova Aida Kh., PhD, docent, Kama Institute (Naberezhnye Chelny, Russia), ANO «Scientific and Research Center for Information in Physics and Technique» (Nizhny Novgorod, Russia), E-mail: aida_khatif@mail.ru

Charnine Mikhail M., PhD, Senior Researcher, FRC CSC of the Russian Academy of Sciences, Moscow, Russia, E-mail: mc@keywen.com

Klokov Alexey A., graduate student, FRC CSC of the Russian Academy of Sciences, Moscow, Russia, E-mail: aaklokov@yandex.ru

Sokolov Evgenii G., graduate student, FRC CSC of the Russian Academy of Sciences, Moscow, Russia, E-mail: evgeny.sokolov@phystech.edu