An Information Retrieval System for Technology Analysis and Forecasting

Nikita Nikitinsky
NAUMEN
Moscow, Russia
nnikitinskij@naumen.ru

Dmitry Ustalov
Ural Federal University
Yekaterinburg, Russia
dmitry.ustalov@urfu.ru

Sergey Shashev
Tsentr Razrabotki
Sevastopol, Russia
favoritefx@mail.ru

***Abstract*—Expert evaluation of grant proposals and research projects is often facilitated by specialized decision support systems, which analyze research and industry trends in a large domain-dependent text corpus. Despite that there exist production-grade technological forecasting systems for English, Russian patent databases and citation indexes had been developed isolated from the global ones. This complicates technology analysis and forecasting in research conducted in Russia. In this paper, we present a scientific information retrieval system designed for the Russian language. The system uses patents, research papers and government contracts for facilitating the expertise process by providing the experts with relevant documents. Comparison of our system with a popular baseline shows promising results.**

I. Introduction

Technological forecast is a prediction of the future characteristics of useful machines, processes, or techniques [1]. People make technological forecasts for many reasons including risk management, market analysis, demand planning, etc. Predicting the future is obviously an exceptionally difficult problem since we may only provide estimates under uncertainty. Since that technological forecasting is a domain-dependent and data-driven kind of activity, it is necessary to provide a context that specifies particular data sources and requirements of a decision making person.

In various government and non-government scientific endowments, invited or employed experts are reviewing incoming grant proposals and deciding whether a given research project should be or should not be awarded with a grant or other kind of benefit. Opinions of such experts may be biased by some reasons: field multidisciplinarity, innovativeness factor, and so forth. Experts use various tools for clarifying their decisions: citation indexes, patent databases, electronic libraries, etc. These data are often closed source and can barely be used openly. Nevertheless, producing new informative features for facilitating the technological forecasting tools may reduce the rate of errors made by decision makers.

In order to facilitate the technology forecasting process, we have created a scientific information retrieval system for the Russian language. Our system uses three data sources (patents, papers and contracts) and exploits word embeddings for extending the user queries, while relying on traditional information retrieval (IR) techniques for the rest of search process. To the best of our knowledge, this is the first attempt to use such query extension techniques and data sources for developing a scientific IR system for the Russian language.

The work, as described in this paper, is focused on the following aspects. Firstly, we will present the architecture of our information retrieval system designed for technological forecasting. Secondly, we will assess the used word embedding technique. Thirdly, a domain expert will evaluate the search quality of our system. Finally, we will compare the system performance with a popular baseline system. As the result, we found that our skip-gram word embedding model performs fair on a comprehensive semantic similarity benchmark for Russian, despite it has been trained on a narrow domain data. Having compared our system with a similar one, we found that it outperforms its in terms of both precision and recall.

The rest of this paper is organized as follows. Section II reviews the related work. Section III defines the problem statement and describes the given data. Section IV presents our scientific information retrieval system (IRS). Section V describes experiments on our system. Section VI shows and discusses the obtained results. Section VII concludes with final remarks and proposes directions for further studies.

II. Related Work

Today, technological forecasting is a multidisciplinary field uniting both quantitative and qualitative methods for recognizing patterns in unstructured data sources and representing such patterns for facilitating the decision making [2]. Techniques for technological forecasting vary from domain to domain, considering different time periods. Despite that there are vast amounts of technological forecasting research, only few attempts have been made for providing a literature review. The most recent review has been conducted by Kang et al. [3] suggesting that traditional data sources for technological forecasting are news materials, patents, research papers and citation databases.

In our short survey, we will focus on two aspects. Firstly, we will review *data collection and analysis* methods used for technological forecasting. Secondly, we will refer to several *publicly available systems* that are often used both in academia and industry for analyzing technological and scientific trends.

A. Data Collection & Analysis

Since that the scientific information retrieval systems are often proprietary, technical details on their implementation are barely available. Nevertheless, most IR systems are based on vector space models, which are proven to be effective in addressing such problems [4]. A good forecasting begins with representative data, hence, in scientific information retrieval, a

ISBN 978-952-68397-0-7 (paperback), ISBN 978-952-68397-1-4 (PDF)

special attention is given to data preprocessing, because it has become highly topical to find better data sources, which either contain valuable insights or might be used in search query extension [5], [6].

Formal results of research and development (R&D) activities are patents, reports, papers, contracts, etc. In 1960s, Garfield & Sher proposed impact factor for evaluating the scientific literature using citation indexes [7]. Recent studies are focused on integrating multiple heterogeneous data sources for producing better output. Patents and research papers are extremely popular data sources in this field of study [8]. Gao et al. used a nearest neighbour classifier for measuring the technology's life cycle stage, which resulted in better forecasting [9]. Kim, Suh & Park proposed a graph-based method for visualizing patent databases by constructing a semantic network from the extracted keywords clustered by the k-means algorithm [10]. Oh et al. combined TF-IDF with WordNet similarity-based clustering for producing bioinformatical forecasts on scientific journals [11]. Woon, Aung & Madnick used TF-IDF and Google distance for predicting trends on the Scopus database [12]. In fact, usage of publications and patents demonstrate robust results, as shown by de Godoy Daiha et al. [13].

In some cases, researchers make use of alternative data sources for improving predictions. Lin et al. proposed using Twitter data for extracting top news, trending topics, active users and top sources to derive a technology trend line [14]. Gök, Waterworth & Shapira showed that materials published on technological companies' websites are a good proxy for technological trend observations [15].

B. Publicly Available Systems

There are several well-known production-grade systems for technological forecasting and related activities (Table I). Those include Questel Orbit [16], Web of Science by Thomson Reuters [17], SciVal by Elsevier [18], Google Patents with Scholar search for English [19], Exactus Expert and Exactus Patent systems by ISA RAS for Russian [20].

According to our analysis, the only available products working with Russian patents are Exactus ones, although these systems do not use government contracts' data.

TABLE I. SYSTEMS FOR TECHNOLOGICAL FORECASTING

	Patents	**Papers**	**Contracts**	**Citations**	**Russian**
Orbit	Yes	No	No	No	No
WoS	No	Yes	No	Yes	No
SciVal	Yes	Yes	No	Yes	No
Google	Yes	Yes	No	Yes	No
Exactus	Yes	Yes	No	No	Yes

III. PROBLEM

Most of currently available systems estimate the state of a research topic given written artifacts. The most widely used input data are patents and different types of publications. However, impact of governmental funding to research areas is often underestimated (Table I).

The experts and decision-makers often need to conduct analysis of research areas, carry out examination of the prior art and provide decision-making with respect to some contract or grant application. This is done in order to prevent plagiarism and conducting a contract work or research, which has already been performed. Furthermore, for experts and decision-makers it is essential to be able to search for certain persons and organizations, experienced in some specified research topics, e.g. in order to find the most suitable contractors for R&D work. Moreover, in terms of conducting prior art search and technology analysis, the researchers as well as decision makers often need to analyze the broader research topic. This problem is sometimes intensified by the fact that decision maker may not be able to compose a required query as he or she might not know in detail the research area of interest.

Our scientific information retrieval system enables the user to conduct search on three types of data: research papers, patents and government contracts with an ability to automatically create context extensions for the initial user query in order to provide a user with opportunity to search for those research topics, which are not familiar for a user.

We have been provided with three following data sources. Firstly, a collection containg $1,119,689$ *invention patents* granted by the Federal Institute of Industrial Property [21], 1984–2014. Secondly, a corpus representing of $884,395$ *research papers* from the Russian Science Citation Index [22], the Russia's largest digital scientific library, 2006–2014. Finally, a subset of $14,375$ Federal Target Programme "Research and Development" *government contracts'* texts provided by the Directorate of Science and Technology Programmes [23], 2005–2014.

IV. SYSTEM DESCRIPTION

We developed a scientific information retrieval for experts and decision-makers, which performs search on Russian patents, research papers and government contracts. The system utilizes classical full-text search method based on the BM25F ranking function [24], but also we use latent semantic analysis and word embeddings for semantic search and user query extension.

A primary idea of the approach in the system is to extend a user query with semantically close terms in order to obtain broader results for the given research topic and then visualize them to let the user see the present trends in the research area.

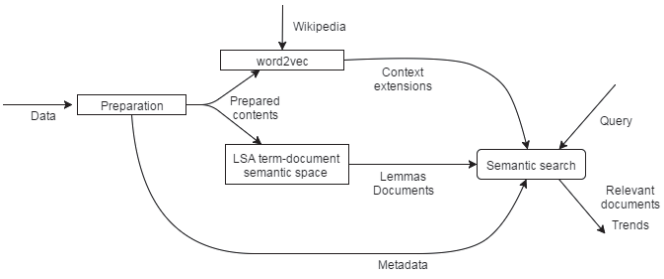

Fig. 1. High-level schema of the system pipeline

The high-level process of building the system is depicted at Fig. 1 and consists of the following steps. First, we prepare the input documents in order to obtain data suitable for other stages of the system. Second, we construct the LSA term-document semantic space. This space is constructed in order to be able

to conduct semantic search: project a given query written in natural language to LSA space in order to retrieve most similar documents and key terms. Then, we built word2vec language model in order to give a user an ability to extend his or her query with the most suitable terms. This approach allows the user to perform search with high recall even for topics, which are not user's areas of expertize. After all, we created the user interface, which allows the user to enter the query, extend it with contextually close terms, set filters and perform the search itself. As the result, the user receives the most relevant documents, persons and organizations. Also, he or she retrieves time series plots for his or her query. More details on every step might be found below.

A. Data Preparation

First, we extract all the documents' metadata, i.e. document types (patent, paper or contract), publishing dates, titles, author names, etc. This is done in order to reach some goals: (1) to be able to set filters for search, e.g. limit publishing date range, (2) to let the user conduct search in various fields separately, e.g. document contents or document title etc., and (3) to construct time series plots for different document types.

Then, we tokenize the contents. During the tokenization process we remove all non-word characters (e.g. punctuation marks) and stop-words, i.e. words having no or almost no sense for analysis, for instance, pronouns and interjections.

Finally, we conduct morphological analysis. In our case, it is composed of lemmatization and part-of-speech tagging. The aim of lemmatization is to reduce inflectional forms of a word to a common base form called lemma. This dramatically reduces the number of unique terms as different forms of one word (e.g. speak, speaks and spoke) are converted into one conventional form (e.g. speak).

The pipeline, consisting of tokenization, stop-words removal and lemmatization is somehow standard preprocessing approach, so we will not go further into detail regarding its advantages.

B. LSA Semantic Space Construction

Then, we create a semantic space using the latent semantic analysis (LSA), which sometimes referred to as latent semantic indexing (LSI). LSA is is a natural language processing technique often used in information retrieval, which analyzes relationships between a set of documents and the terms they contain [25]. The general assumption of LSA is that words with similar meaning tend to occur in similar contexts. We have chosen LSA since this approach is fast compared to other methods, robust as it only involves decomposing the term-document matrix, can easily be trained on big data sets and can partially handle homonymy. For us, the speed and consistency in terms of sensitivity to starting conditions played a major role, since we had a big data set consisted of various types of documents to be processed.

In this technique, a weighted term-document matrix is constructed, where rows represent unique words, and columns represent documents. The matrix is built using well-known mathematical technique called singular value decomposition (SVD) in order to reduce the number of rows. SVD is computed as follows:

$$\mathrm{M} = U\Sigma V^* \tag{1}$$

where M is $m \times n$ matrix whose entries come from some field K, U is $m \times m$ matrix, Σ is $m \times n$ diagonal matrix with non-negative real numbers on the diagonal and V^* is an $n \times n$ unitary matrix over K.

We have chosen the LSA with Log Entropy weighting function, because they work well in many practical studies [26]. Particularly, each cell a_{ij} of a term-document matrix A is computed as follows:

$$\mathrm{p}_{ij} = \frac{tf_{ij}}{gf_i}, \tag{2}$$

$$\mathrm{g}_i = 1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}, \tag{3}$$

$$\mathrm{a}_{ij} = g_i \times \log(tf_{ij} + 1), \tag{4}$$

where n is total number of documents, g_i is the global weight, tf_{ij} is the number of occurrences of term i in document j, and gf_i is the total number of times the term i occurs in the corpus.

Having constructed the LSA term-document space, we extract key terms by computing cosine similarity between documents and terms. For each document, we consider terms having similarity more than 0.8 key terms. In this sense, our approach is similar to one shown by L'Huillier et al. [27].

C. Creating a word2vec model

After extracting the key terms, we train a word2vec model on the collection of patents, papers, contracts combined with the Russian Wikipedia.

Word2vec is a tool implementing two shallow neural network architectures, namely skip-gram and continuous bag-of-words (CBOW), used for computing vector representations of words [28]. The general assumption of word2vec is the same as that of LSA: similar words tend to occur in similar contexts. In our implementation, we use the skip-gram architecture, as it works better for non-frequent words than CBOW, another popular shallow neural network architecture. Here, given a sequence of training words $w_1, w_2, w_3, \ldots, w_T$, the training objective of the skip-gram model is to maximize the log probability

$$\frac{1}{T}\sum_{t=1}^{T}[\sum_{j=-k}^{k} \log p(w_{t+j}|w_t)] \tag{5}$$

where k is the size of the training context.

D. User Interface

Search process is being conducted as follows. Initially, a user composes a simple query on topic he or she want to analyse, e.g "polymer". Then, the system extends the query to obtain broader results for the topic. Text of the user query is extended with contextually similar words from the trained word2vec model and then searched in the LSA semantic space in order to retrieve the most relevant documents. The extended query might be corrected by a user in order to exclude the

words that user considers irrelevant for the given query. Example of the above mentioned query extension approach for the word "polymer" is depicted at Fig. 2. Particularly, the system extracted top-20 nearest neighbors for the query "polymer" in the Russian language: *polymeric, oligomer, copolymer, siloxane, macromolecular, monomer, macromolecule, oligomeric, supramolecular, polyether, organosilicon, elastomer*, etc.

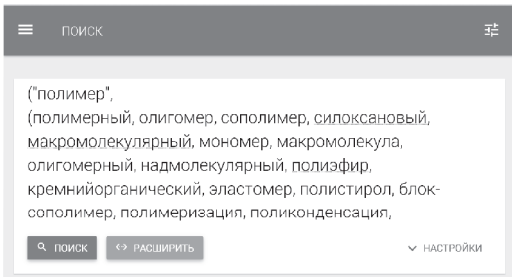

Fig. 2. Example of semantic context extensions in Russian

This approach makes it possible for user to conduct an analysis of a research topic and control the degree of breadth of the topic with context extensions. Since common use cases of the system include analysis of certain research area in order to research the prior art in the area, decision making with respect to some contract and search of certain persons and organizations, experienced in some specified research topics, users of our system require us to provide the paper, patent and contract data separately. This information is used by them in order to facilitate decision making, for example, with reference to the particular grant application. Nevertheless, the system has an option to retrieve paper, patent and contract data altogether ranked by relevance to the query. In our further experiments we use only separate patent, paper and contract data retrieval.

V. Experiments

In this section, we evaluate the performance of our IRS system. We prepare data and conduct the following experiments:

1) evaluation of the word2vec model in order to estimate its ability to produce context extensions of decent quality,
2) evaluation of IRS on a test set in order to verify the ability of context extensions to improve the results of information retrieval in comparison with search based on an initial query,
3) evaluation of IRS against a Baseline search system in terms of precision and recall.

A. Evaluation of word2vec

In order to evaluate our word2vec model (Section IV-C), we used the AE2 and RT test sets from the Russian Semantic Similarity Evaluation workshop (RUSSE) [29]. The AE2 dataset, based on a cognitive experiment, measures how well a system estimates association between words. The RT dataset, based on a popular lexical ontology for Russian, evaluates the ability of a system to identify synonyms and hypernyms. As in the original RUSSE study, we used average precision as the performance measure:

$$AveP = \frac{\sum_r P@r}{R}, \tag{6}$$

where r is the rank of each relevant document, R is the total number of relevant documents, and $P@r$ is the precision of the top-r retrieved documents [30].

In RUSSE, the systems have been trained on much larger document collections representing common lexis, e.g. the whole Russian Wikipedia or vast digital libraries. For instance, the RUSSE organizers report that the highest ranked AE2 system shows average precision of 0.9849 and the RT system demonstrates 0.9589. Though our word2vec model has been trained on documents belonging to a specific domain, i.e. research papers, patents and contracts, it ranked 11th out of 19 best systems on both test sets (Table II).

TABLE II. Evaluating the word2vec model on RUSSE

	AE2	**RT**
$AveP$	0.9107	0.7639
Rank	11	11

Accordingly, we suppose that the quality of context extensions produced by the word2vec model is reasonably fair.

B. IRS evaluation

In this subsection, to evaluate IRS we selected 10 different research areas, which included polymer technologies, medical technologies and information technologies. Research areas were randomly selected in accordance with Federal Target Programme "Research and Development" priority directions, which are (1) life sciences, (2) nanotechnologies and new materials, (3) information and Telecommunication technologies; (4) rational use of natural resources, (5) energy efficiency, energy saving and nuclear energy. Then, for every research area experts created a test set consisting of 100 documents for each data type (patents, papers, contracts). The test set was assembled from the most relevant documents for the research area issued in 2008.

To show an advantage of using context extensions for retrieving more relevant documents for a research area, we retrieved documents in 5 different ways:

1) Using one-word direct query search, i.e. without applying context extensions. We used one most important word for a research area, e.g. for polymer technologies area we used the query "polymer" in Russian.
2) Applying context extension to a query and then removing the initial query. This is done in order to show the possible performance of the context extension itself.
3) Applying context extension to a query, removing the initial query and manually editing the extension in order to delete the most inappropriate terms. In our experiments, we manually removed from 1 up to 4 out of 20 possible extension terms.
4) Applying context extension to an initial query and using it for search along with the initial query.
5) Applying context extension to a query, then manually editing the extension in order to delete the most

inappropriate terms and using the extension for search along with the initial query. In our experiments, we manually removed from 1 up to 4 out of 20 possible extension terms.

In this study, we first compared the most relevant 100 documents retrieved by IRS to the expert data set and then evaluated the precision of information retrieval on the most relevant 1000 documents retrieved by IRS. We computed the mean precision for the retrieved results at first 100 retrieved documents (P@100). We used the classic precision measure for information retrieval, which is defined as the fraction of retrieved documents that are relevant to the query [4]:

$$P = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}, \tag{7}$$

where D_{rel} is number of relevant documents and D_{retr} is number of retrieved documents.

For the precision-recall plot, we needed to compute recall, which is expressed as the fraction of the documents that are relevant to the query that are successfully retrieved:

$$R = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|}. \tag{8}$$

It is clear that recall of 100% can be achieved by returning all documents in response to any query. Therefore, we use this measure only for precision-recall plots (see Fig. 3, 4, 5).

In this study, we also make use of Receiver Operating Characteristic (ROC) curves. ROC curves are typically used in information retrieval and binary classification to study the output of the classifier. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR). In terms of information retrieval, TPR is equal to recall and FPR can be expressed as:

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

where FP is the number of false positive retrieved results and TN is the number of true negative retrieved results

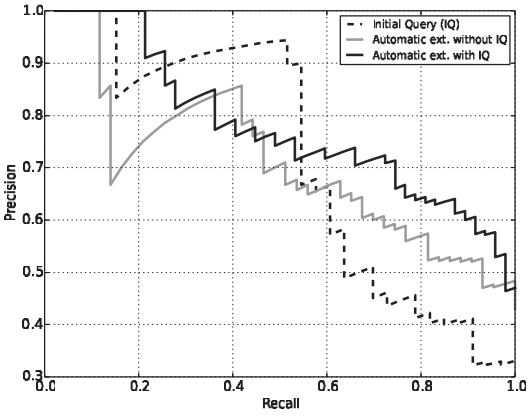

Fig. 3. Precision-recall plot for patents on P@100

In this study, we also make use of ROC curves()

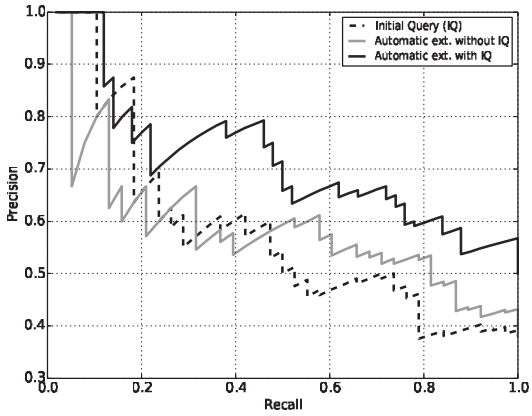

Fig. 4. Precision-recall plot for papers on P@100

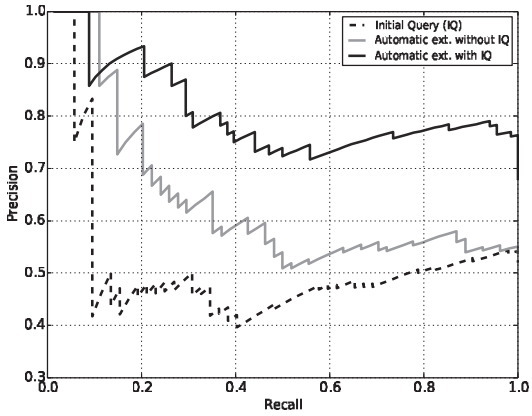

Fig. 5. Precision-recall plot for contracts on P@100

According to the results from Table III, we may see that IRS shows higher precision at first 100 retrieved documents (P@100) for the cases when we applied context extensions (both automatically created and manually edited): 56.1% and 57.9% respectively. A small difference (1.8%) between the automatically created and manually edited context extensions indicates the high relevance of automatic context extensions.

TABLE III. P@100 UNDER DIFFERENT SEARCH OPTIONS

	Patents	**Papers**	**Contracts**	**Mean**
Initial query (IQ)	32.5	38.2	52.4	41.0
Automatic ext. without IQ	45.1	38.0	54.3	45.8
Manually edited ext. without IQ	49.1	37.3	49.8	45.4
Automatic ext. with IQ	49.9	50.3	68.1	56.1
Manually edited ext. with IQ	51.2	50.2	72.2	57.9

TABLE IV. P@1000 UNDER DIFFERENT SEARCH OPTIONS

	Patents	**Papers**	**Contracts**	**Mean**
Initial query (IQ)	95.8	89.4	85.3	90.1
Automatic ext. without IQ	97.1	92.4	92.0	93.8
Manually edited ext. without IQ	98.0	95.2	95.4	96.2
Automatic ext. with IQ	100.0	99.8	98.1	99.3
Manually edited ext. with IQ	100.0	100.0	100.0	100.0

When it comes to the mean precision on the first 1000 retrieved results (P@1000) with reference to the test set, we

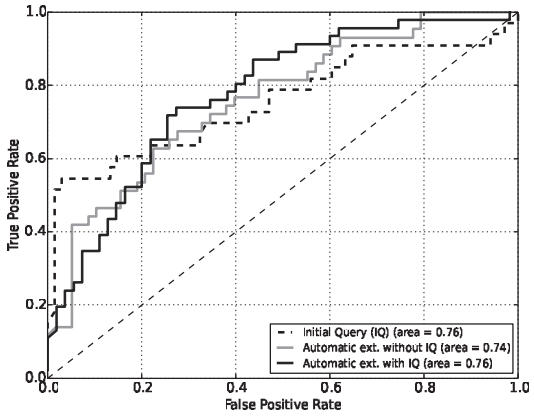

Fig. 6. ROC-plot for patents on $P@100$

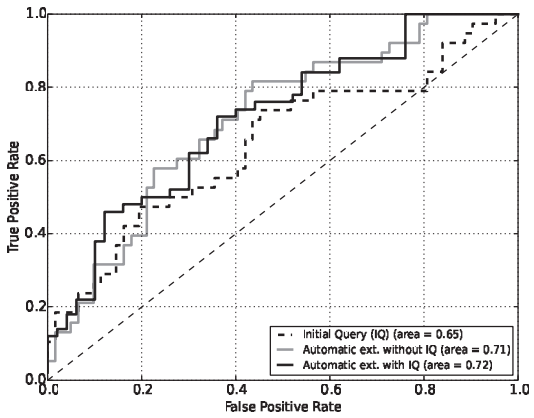

Fig. 7. ROC-plot for patents on $P@100$

may see the following results: initial query with automatically created context extensions along with manually edited context extensions showed the average precision of 99.3% and 100% respectively, while initial one-word query achieved only 90.1% (see Table IV).

C. Comparison with Baseline

We conducted tests on a baseline search system. We chose the same patent database (FIPS: Inventions) and year (2008) for both the baseline and IRS. We took the same one-word queries from 5 different research areas and performed the following tests:

1) Direct one-word query search, i.e. without applying context extensions, for patents in IRS (IRS DS) and in the baseline search system (Baseline DS).
2) Applying context extensions for the initial one-word query for IRS (IRS CE). In order to retrieve the most relevant results and limit the number of retrieved documents, we restricted IRS to retrieve the documents containing the initial query or at least 25% of all automatic context extensions.

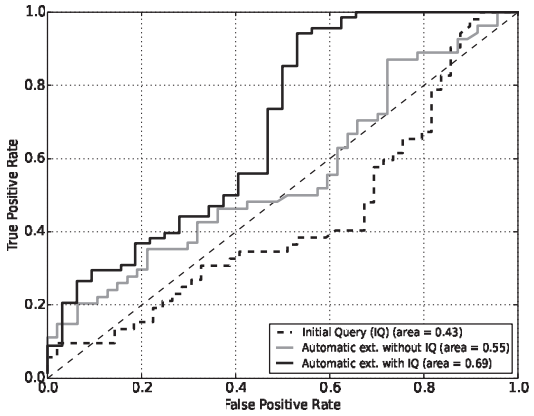

Fig. 8. ROC-plot for contracts on $P@100$

To measure relative change in recall, we assumed the baseline system to have 100% recall.

As the result, we found out that IRS has higher recall (+4.3% in average if we consider the baseline as 100%) when we utilized direct one-word query search. At the same time, the mean increase in recall when we applied context extensions was +88.9%. See Table V and Fig. 9 for details.

TABLE V. COMPARISON OF IRS WITH BASELINE IN TERMS OF RECALL

	Baseline DS	**IRS DS**	**IRS CE**
Polymer	1913	2007	4336
Medicine	3556	3801	3969
Aviation	179	190	360
Radio Physics	16	16	17
Petroleum	1025	1062	3061

TABLE VI. COMPARISON OF IRS WITH BASELINE IN TERMS OF PRECISION ($P@10$)

	Baseline DS	**IRS DS**	**IRS CE**
Polymer	30	30	50
Medicine	20	20	30
Aviation	20	30	40
Radio Physics	90	80	90
Petroleum	30	40	40

We measured precision for Baseline and IRS at first 10 retrieved documents ($P@10$) compared to 10 patents sampled by experts as most relevant for every research area under study. The mean precision for the Baseline amounted to 38% while for IRS it was 40% for direct one-word query search and 50% for search with initial query plus automatically created context extensions. See Table VI and Fig. 10 for details.

VI. RESULTS AND DISCUSSION

In this study, we introduced an approach to research topic analysis for decision makers and experts, based on extending a user query with semantically close terms in order to obtain broader results for the research topic and the information retrieval system (IRS) exploiting this approach.

The approach is robust and shows better recall and precision (+4.3% recall and +2% $P@10$ precision improvement) compared to the Baseline search system when direct one-word

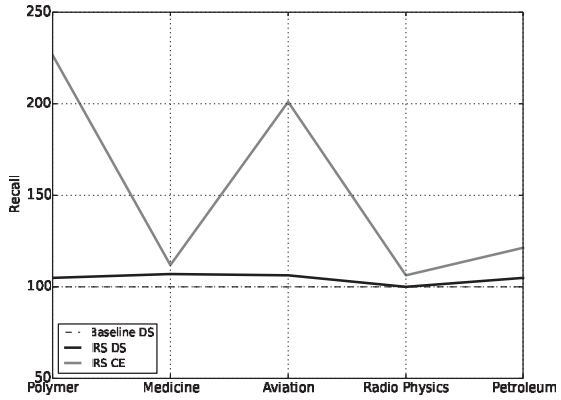

Fig. 9. Comparison of IRS with baseline on 5 topics in terms of recall

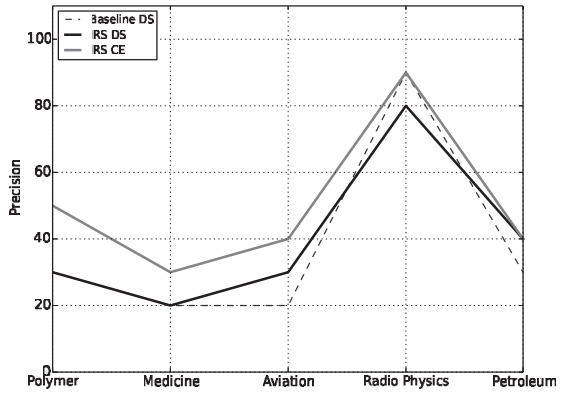

Fig. 10. Comparison of IRS with baseline on 5 topics in terms of precision ($P@10$)

query search is used. Employment of automatically created context extensions improved the results even more: up to $+226\%$ (recall) and $+12\%$ ($P@10$ precision).

Evaluating IRS on a test set, we obtain higher precision at first 100 retrieved documents ($P@100$) on test data set when applied automatically created context extensions to the initial one-word query: the mean precision was 56.1%. Manual correction of the context extension (we deleted from 1 to 4 most irrelevant terms out of 20 possible terms in extension) improved the mean precision to 57.9%. The small difference in the mean precision between the automatically created and manually edited context extensions may indicate high relevance of automatically created context extensions.

The precision at first 1000 retrieved documents ($P@1000$) shows us that applying context extension to the initial one-word query may improve the precision up to 10% (i.e. from 90.1% with initial query search to 99.3% when adding context extension to initial one-word query and to 100% with manual correction of the context extension applied to the initial query).

The word2vec model, which we used to create context extensions, was evaluated on two test sets from the Russian Semantic Similarity Evaluation benchmark (RUSSE). Our model showed the average precision of 0.7639 and 0.9107 on the test sets RT and AE2, correspondingly. These results rank our model 11th out of 19 best systems on both test sets.

Certainly, our study has limitations. For instance, we used a limited amount of data, especially for patents and contracts. Adding utility models' patents data and contracts for other Federal Programmes to the word2vec model may improve the quality of the context extensions to achieve better information retrieval results in terms of relevance.

VII. Conclusion and Future Research

To conclude, we would like to say that in terms of information retrieval our ISR system show acceptable performance and may be utilized for technology analysis and forecasting.

According to the results of our study, IRS shows higher recall and precision compared to the Baseline search system. At the same time, while evaluating the performance of IRS under different types of queries we found out that the best performance is achieved when applying automatically created and manually edited context extensions along with the initial one-word query. The word2vec model we utilized for creating context extensions demonstrated high performance (ranked 11th out of 19 best systems) when evaluated on test sets from the Russian Semantic Similarity Evaluation workshop (RUSSE).

For a future study, we suggest:

- using more diverse data (e.g. all types of patents, including utility models; and more contracts' data for other Federal Targeted Programmes);
- adding data sources in other languages (primarily, English) to be able to study how the system will work with other languages;
- creating an automatic research area classifier to improve relevance of information retrieval.

We also look forward to applying IRS to analysis of the possible impact of government contracts on technology forecasting.

Acknowledgment

We would like to acknowledge the hard work and commitment from Stepan Kamentsev and Alexey Korobeynikov throughout this study.

A special thank you goes to the Directorate of Science and Technology Programmes and personally to Andrey N. Petrov for providing us with government contracts' data for this study. We are also grateful to the anonymous referees who offered very useful comments on the present paper.

Research reported in this publication was supported by Ministry of Education and Science of the Russian Federation under contract number 14.579.21.0091.

REFERENCES

[1] J.P. Martino, *Technological Forecasting for Decision Making*. McGraw-Hill, Inc., 1993.

[2] R.J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2013.

[3] D. Kang, W. Jang, H. Lee, H.J. No, "A review on technology forecasting methods and their application area", *World Academy of Science, Engineering and Technology*, vol.7, Apr. 2013, pp. 394-398.

[4] C.D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press. 2008.

[5] C. Carpineto, G. Romano, "A Survey of Automatic Query Expansion in Information Retrieval", *ACM Computing Surveys*, vol.44(1), Jan. 2012, pp. 1:1-1:50.

[6] W. Giänzel, "Bibliometrics-aided retrieval: where information retrieval meets scientometrics", *Scientometrics*, vol.102(3), Mar. 2015, pp. 2215-2222.

[7] E. Garfield, I.H. Sher, "New Factors in the Evaluation of Scientific Literature Through Citation Indexing", *American Documentation*, vol.14(3), Jul. 1963, p. 195-201.

[8] T.U. Daim, G. Rueda, H. Martin, P. Gerdsri, "Forecasting emerging technologies: use of bibliometrics and patent analysis", *Technological Forecasting and Social Change*, vol.73(8), Oct. 2006, pp. 981-1012.

[9] L. Gao, A.L. Porter, J. Wang, S. Fang, X. Zhang, T. Ma, W. Wang, L. Huang, "Technology life cycle analysis method based on patent documents", *Technological Forecasting and Social Change*, vol.80(3), Mar. 2013, pp. 398-407.

[10] Y.G. Kim, J.H. Suh, S.C. Park, "Visualization of patent analysis for emerging technology", *Expert Systems with Applications*, vol.34(3), Apr. 2008, pp. 1804-1812.

[11] K.J. Oh, C.-G. Lim, S.S. Kim, H.-J. Choi, "Research trend analysis using word similarities and clusters", *International Journal of Multimedia and Ubiquitous Engineering*, vol.8, Jan. 2013, pp. 185-196.

[12] W.L. Woon, Z. Aung, S. Madnick, "Forecasting and visualization of renewable energy technologies using keyword taxonomies", *in 2nd International Workshop on Data Analytics for Renewable Energy Integration*, Sep. 2014, pp. 122-136.

[13] K. de Godoy Daiha, R. Angeli, S.D. de Oliveira, R.V. Almeida, "Are lipases still important biocatalysts? A study of scientific publications and patents for technological forecasting", *PLOS ONE*, vol.10(4), Jun. 2015, pp. 1-20.

[14] Y.-C. Lin, P. Yang, W.-T. Hsieh, S.T. Seng-Cho, "Technology trend analysis tool using twitter as a source", *International Journal of Information Technology & Computer Science*, vol.6, Dec. 2012, pp. 69-74.

[15] A. Gök, A., Waterworth, P. Shapira, "Use of web mining in studying innovation", *Scientometrics*, vol.102(1), Jan. 2015, pp. 653-671.

[16] Questel - Innovation, Invention, Patent, Licensing, Web: http://www.questel.com/index.php/en/.

[17] Web of Science, Web: http://webofknowledge.com/.

[18] SciVal - Welcome to SciVal, Web: https://www.scival.com/.

[19] Google Patents, Web: https://patents.google.com/.

[20] ISA RAS website - Homepage, Web: http://www.isa.ru/index.php?lang=en.

[21] Federal Institue of Industrial Property, Web: http://www1.fips.ru/wps/wcm/connect/content_en/en/.

[22] Russian Science Citation Index, Web: http://elibrary.ru/.

[23] Directorate of Science and Technology Programmes, Web: http://www.fcntp.ru/.

[24] J.R. Perez-Aguera, J. Arrovo, J. Greenberg, J.P. Iglesias, V. Fresno, "Using BM25F for semantic search", *in Proceedings of the 3rd International Semantic Search Workshop SEMSEARCH '10*, Apr. 2010.

[25] S. Deerwester, S. T. Dumais, G. W. Furnas, T.K. Landauer, R. Harshman., "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, 41 (6), 1990, pp. 391-407.

[26] T.Landauer, D.S. McNamara, S.Dennis and W. Kintsch., *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007.

[27] G. L'Huillier, A. Hevia, R. Weber, S. Rios, "Latent Semantic Analysis and Keyword Extraction for Phishing Classification", *2010 IEEE International Conference on Intelligence and Security Informatics*, May 2010, pp. 129-131.

[28] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean., "Distributed representations of words and phrases and their compositionality", *Neural Information Processing Systems 2013*, Dec. 2013, pp. 3111-3119.

[29] A. Panchenko, N.V. Loukachevitch, D. Ustalov, D. Paperno, C.M. Meyer, N. Konstantinova, "RUSSE: The First Workshop on Russian Semantic Similarity", *in Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*, vol.2, May 2015, pp. 89-105.

[30] L. Liu and M.T. Özsu, *Encyclopedia of Database Systems*, New York: Springer US, 2009.