Sci-Search: Academic Search and Analysis System Based on Keyphrases

Svetlana Popova[1,2], Ivan Khodyrev[3,4], Artem Egorov[2], Stepan Logvin[2], Sergey Gulyaev[2], Maria Karpova[2], and Dmitry Mouromtsev[5]

[1] Saint-Petersburg State University, Saint-Petersburg Russia
svp@list.ru
[2] Saint-Petersburg State Polytechnic University, Saint-Petersburg Russia
ARTEGO@fuit.spbstu.ru
[3] Saint-Petersburg State Electrotechnical University, Saint-Petersburg Russia
[4] VISmart, Saint-Petersburg Russia
kivan.mih@gmail.com
[5] Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint-Petersburg Russia
d.muromtsev@gmail.com

Abstract. Structured data representation allows saving much time during relevant information search and gives a useful view on a domain. It allows researchers to find relevant publications faster and getting insights about tendencies and dynamics of a particular scientific domain as well as finding emerging topics. Sorted lists of search results provided by the popular search engines are not suitable for such a task. In this paper we demonstrate a demo version of a search engine working with abstracts of scientific articles and providing structured representation of information to the user. Keyphrases are used as the basis for processing algorithms and representation. Some algorithm details are described in the paper. A number of test requests and their results are discussed.

Keywords: Representing search results, academic search engine, keyphrase extraction, clustering, indexing, informational retrieval.

1 Introduction

Internet could be considered as a dynamic source of information, which is constantly updating. Possibility to gather and process new data could be used to obtain new knowledge. In the field of science it is especially important. To work efficiently researcher needs to have an overview of the current state of the art in his field, which allows him to choose the strategy of research and save time. However to achieve this goal to have a good searching methods of relevant documents is not enough. In addition, it is important to represent search results in a suitable form for the researcher's needs. In times when the number of online scientific articles grows very fast and a search query usually results in a dozens of relevant documents, the structured presentation of results comes into first plane. In this paper we present an academic search

P. Klinov and D. Mouromtsev (Eds.): KESW 2013, CCIS 394, pp. 281–288, 2013.

system (http://sci-search.uni.me), which helps in searching scientific articles and presenting the state of a scientific domain.

2 Related Work

Today there exists a number of search engines related to databases of scientific publications: Microsoft Academic Search[1], Scirus[2], CiteSeerX[3], Google Scholar[4] and others. Traditionally, a result of searching requests in these systems is presented as a list of documents. These systems also usually allow facet search. Most of the systems use relevant keyphrases, by clicking on which ones can obtain quite a long list of documents. However, such a list doesn't reflect dynamics of a particular domain and doesn't allow skipping a group of thematically close documents, which currently have no interest. Another problem is a small number of documents for some specific domains making it hard to find information. In this case a list representation would shuffle relevant documents with irrelevant which further confuses researcher. The main goal for our system is to avoid mentioned problems.

Sci-Search search engine is based on active usage of keyphrases. Comparing to lists of words, keyphrases allow not only extract important words, but also reflect the context of their usage. In [1] a search interface using keyphrases for document representation is proposed. Authors show that such presentation is more convenient for end-users. In a search engine Keyphind (Phind) [2] keyphrases were used for indexing. Search results are represents as a list of keyphrases and each keyphrase is associated with a group of documents and other phrases. In [3 - 5] a problem of the searching results clustering is addressed, without focusing on academic search. These approaches use ranked query result lists from search engines like Google and they build keyphrases using only titles and snippets.

In Sci-Search system we integrate keyphrase-based clustering of obtained search results with keyphrase annotations of clusters and documents. To generate keyphrases we use titles of publications and abstracts, based on assumption that abstracts reflect main theme of an article. We extract keywords that are popular in each year and for the whole period for each query. For the most popular keywords, we depict dynamics of their popularity for a given period. We assume that such presentation of search results could help scientists in domain analysis and in search of interesting papers.

3 System Description

The user interface (Fig. 1) contains a number of tabs containing different points of view on the search query results. A list of related keyphrases is shown for every retrieved document. We present results from the "Analytic info" and "Diagrams" tabs,

[1] http://academic.research.microsoft.com
[2] http://www.scirus.com
[3] http://citeseer.ist.psu.edu/index
[4] http://scholar.google.ru

because they contain demonstrative representation of a query and help user in domain analysis. The “Diagrams” tab illustrates popularity dynamics of twenty most significant keyphrases in documents obtained by request. The “Analytic info” tab returns useful information for analysis of domain area, such information is grouped by year and includes most popular keywords and a set of clusters, annotated with keyphrases. Annotation to each cluster contains two types of keyphrases: phrases occurred in most documents of a cluster and phrases which mostly characterize current cluster and not other clusters. First type helps in identification of thematically interesting groups of documents. Second one allows not to miss more rare but still significant themes. As information is grouped by year, it allows estimation of general dynamics in the domain of search query.

In Sci-Search system all the data, gathered by crawler, is automatically annotated with keyphrases, which are then used in indexing. Clustering algorithm is used on a stage of query results processing.

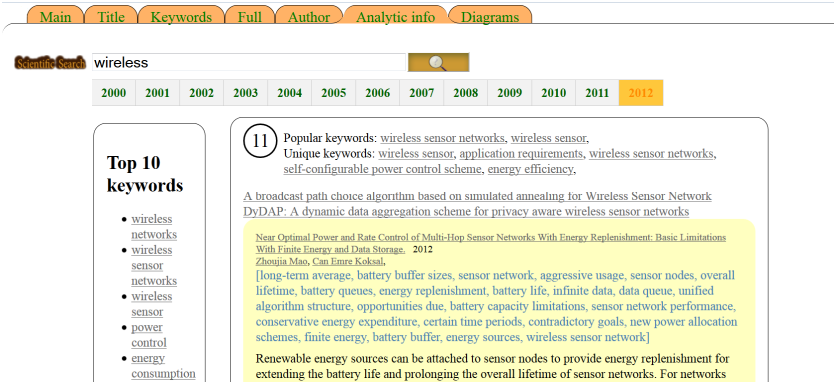

Fig. 1. User interface Sci-Search System

3.1 Extraction of Keyphrases

Nouns and adjectives are the most useful parts of speech to solve a keyphrase extraction problem [6-10]. Additionally we have noticed that in abstracts of scientific articles most multiword sequences of nouns and objectives are strongly related to the general theme of an article. In our research [11] we proposed the simple approach for keyphrase extraction from abstracts of scientific articles having relatively high results comparing to the state of the art. We discuss it in this part. Our approach is based on empiric observation, that among possible keyphrase sequences of nouns and adjectives consisted from only one word, the number of correct keyphrases is much less than in multiword keyphrase. In the same time, the number of one word sequences usually greatly exceeds the number of multiword sequences. Thus, cutting one word sequences allows improving the quality of keyphrases extraction and our experiments proves it. We reduced keyphrase extraction task to the extraction of sequences of nouns and adjectives with length from 2 to 5. Phrases are extracted during the process of consecutive word reading from a document. When a noun or an adjective occurs,

the phrase building begins. It finishes with the occurrence of a word being not noun or adjective, or a punctuation mark, or when phrase length reaches 5 words. In any of these cases the built phrase is added to the list of keyphrases (if the phrase has the length of at least 2 words) and a parser waits for a next occurrence of noun or adjective. For part-of-speech tagging we used Stanford POS tagging tool[5]. All our experiments in [11] were conducted on Inspec dataset [6], [7], [10] containing scientific abstracts and a golden standard of keyphrases for each document (keyphrases extracted by experts). We compared results of our algorithm with the golden standard. F-measure to evaluate the quality of our results was used [6]. For proposed algorithm F-measure=0.40 improves results from [6, 7] and is computationally more simple. Our approach also doesn't need a machine learning as, for example, Kea used in [2]. Table 1 contains examples of abstracts and automatically extracted keyphrases.

Table 1. Examples of automatic extracted keyphrases

Abstracts	Keyphrases
A frequency-tunable and switched beam antenna is proposed. Variable capacitances are loaded to monopole antenna elements and switch the radiation beam as well as the frequency band. Numerical results of the frequency band and beam pattern of the proposed antenna are shown with varying the value of variable capacitances.	*beam antenna, radiation beam, frequency band, antenna elements, beam pattern, variable capacitances, disaster-resilient wireless*
In recent ten years, wireless sensor network technology has a rapid development. After a brief introduction of the wireless sensor network, some main research results of energy conservation and node deployment is provided. Then the applications of WSN in the medical health, environment and agriculture, intelligent home furnishing and building, military, space and marine exploration are outlined. In addition, we analyze the advantage of WSN in these areas. Finally, we summarize the main factors that affect the applications of wireless sensor network	*wireless sensor network technology, wireless sensor, brief introduction, medical health, node deployment, energy conservation, marine exploration, intelligent home furnishing, main research results, rapid development, main factors, wireless sensor network*

In Sci-Search system we are using additional list of stop-phrases, such as: «previous approaches», «other hand», «last works», etc. It was built by an expert using frequency statistics of most used phrases in scientific abstracts. We didn't make a difference between phrases being identical after stemming or containing different word order in a phrase. For stemming we divided phrases into single words and used a Porter stemmer[6].

3.2 Clustering

For clustering we used an algorithm based on k-means variation. Each document was represented as a set of keyphrases. On first stage, the cluster centroids were built.

[5] http://nlp.stanford.edu/software/tagger.shtml
[6] http://tartarus.org/martin/PorterStemmer

Each centroid is a set of keyphrases. At first algorithm runs through all documents and if a set of keyphrases doesn't have an intersection with other centroids with more than one phrase, then this set of keyphrases becomes centroid itself. When the first generation of centroids is built, similar of each document to each centroid are calculated and documents are distributed among centroids according to the most similarity. We have used Jaccard index for similarity calculation:

$$\text{Jaccard}= \frac{A \cap D}{A \cup D}, \tag{1}$$

where A — is a set of keyphrases in a centroid, D — set of keyphrases which represent document. After documents are distributed among clusters, centroid recalculation takes place. New centroids are built as sets of keyphrases, which occurred not less than in 2/3 documents of a cluster. If centroid contains less than 2 keyphrases, it is deleted. When new centroids are built, documents are distributed among clusters with new centroids. Such process is repeated 5 times. Clusters obtained on last iteration are considered as a result.

3.3 Annotation of Clusters with Keyphrases

We used two types of annotations for clusters, which are separately shown to a user. First type ("Popular keywords") reflects main theme of a cluster. To form it, from all keyphrases not more than five most frequent phrases are selected. Each of selected phrases had to occur in more than a half of the documents of a cluster. If a cluster is rather small and there are no phrases fully occured in other documents, then the cluster is represented by a zero number of keyphrases. Second type of annotations ("Unique keywords") should allow finding unique documents characterizing current cluster but not any others. Phrases were selected by weight calculated using formula: N/R where N is a number of documents with that phrase in a cluster and R is a number of documents with that phrase in other clusters.

4 System Demonstration

To obtain data for a test demonstration, the crawler was written gathering data for the theme "the theory of automatic control". The source of data is DBLP[7]. Abstracts of publications were taken from sites of electronic libraries, if a link on DBLP page were provided for a particular paper. For the system demonstration more than 13,000 articles were extracted and processed. We show examples of searching results on tabs "Analytic info" and "Diagrams". Two queries were processed and the results are depicted in Table 2 and Fig. 2 for queries "wireless" and "robot". In Table 2 data about clusters with two types of keywords is presented. First type is called "Popular keywords" and it reflects frequent keyphrases in a cluster, second type, "Unique

[7] http://www.dblp.org

keywords", reflects keywords, which are frequent for this particular cluster and rare for other clusters. Top 10 keyphrases in 2012 year for queries: "wireless" (wireless networks, wireless sensor networks, wireless sensor, power control, energy consumption, interference model, topology control, link scheduling, optimal control, topology control algorithm) and "robot" (mobile robot, humanoid robot, robotic systems, robot manipulators, multi-robot system, robot arm, real robot, real time, robot hand, energy consumption). Fig. 2 depict examples from "Diagram" tab containing usage dynamics of 20 the most popular keywords, related to the query depending on a year and a percentage of documents where they are used among other documents. Fig. 3 represents documents from two clusters for query "wireless" in 2012 year.

Table 2. Clusters and cluster's keyphrases for queries "wireless" and "robot" in 2012 year

Query: "wirelss"	**Query: "robot"**
Cluster size and cluster's keyphrases	*Cluster's keyphrases*
Cluster size: 11 **Popular keywords**: wireless sensor networks, wireless sensor. **Unique keywords**: wireless sensor, energy efficiency, wireless sensor networks, self-configurable power control scheme, application requirements.	Cluster size: 10 **Popular keywords**: mobile robot. **Unique keywords**: fuse different directional identification results, directional information, mobile robot, bandwidth allocation, results identification.
Cluster size: 4 **Popular keywords**: interference model, link scheduling, individual nodes, wireless networks, important goals. **Unique keywords**: stochastic arrival processes, maximum feasible traffic, interference model, overall network topology, important goals.	Cluster size: 5 **Popular keywords**: robotic systems. **Unique keywords**: power plants, scientific knowledge, hierarchical fashion, promising applications, new actions.
Cluster size: 3 **Popular keywords**: energy consumption, network lifetime. **Unique keywords**: local information, multi-hop communication, low-delay route, initial topology, interference reduction.	Cluster size: 5 **Popular keywords**: humanoid robot. **Unique keywords**: compensatory head/eye movements, modular framework, low throughput, brain-computer interfaces, simple form.
Cluster size: 3 **Popular keywords**: wireless networks, power control, link rates. **Unique keywords**: link rates, weighted max-min rate fairness problem, profit maximization, rate max-min fairness, piecewise linear link rate.	Cluster size: 4 **Popular keywords**: energy consumption. **Unique keywords**: energy consumption, stochastic force fields, average power consumption, energy consumption model, angular velocity.

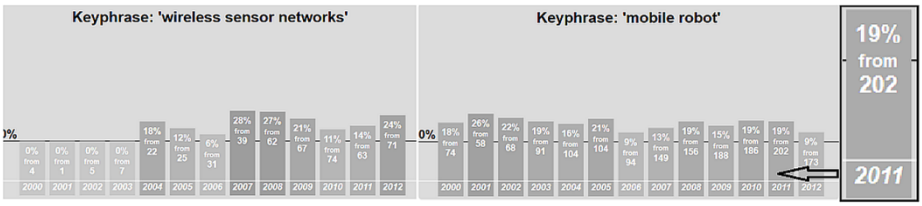

Fig. 2. Dynamics of keyphrases "wireless sensor networks" and "mobile robot" for queries "wireless" and "robot" since 2000 in "Diagrams" tab. Each column contains information about the percentage of documents which contain the phrase to the number of found documents for a particular year.

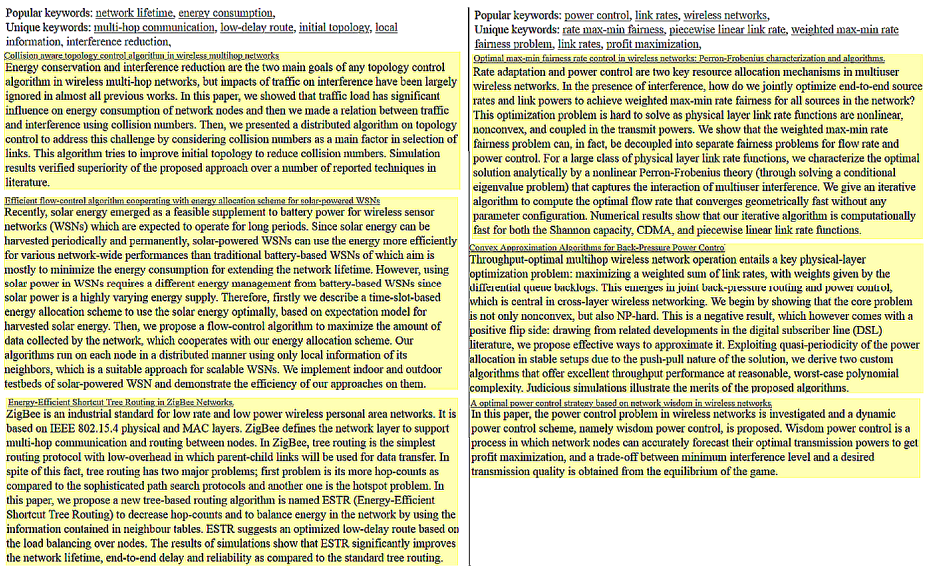

Fig. 3. Contents of two clusters for the query “wireless” in 2012 year (size of each cluster is 3 documents)

5 Discussion and Conclusion

Data collected and processed in the system gives an insight on what happened in a particular scientific domain during some period. Such information could be useful in educational and academic purposes when somebody wants to understand state of the art in a particular area and its dynamics. Formed clusters, described in tables, allow to group thematically close documents reducing a task o scientific domain search task to the search of useful cluster. Cluster annotations allow evaluating related specific themes. Main themes of a search query domain could be extracted using 10 most popular keywords, built for a particular year. Diagrams allow identifying frequently used keyphrases, depending on a year and a percentage of documents where it is used.

Now there are some limitations of the system demo-version: Sci-Search contains information about 13000 papers taken from DBLP and related to the “theory of automatic control” domain. We are planning to increase paper coverage greatly. Now such a small size of a database is a reason of a small number of documents resulted in a query. Cluster forming algorithm dealing with a small number of documents could lead to a poor quality of clustering. However, Pic. 3 and a manual look through the list of clusters in system shows their adequacy, and we assume that it shows perceptiveness of the proposed approach.

Another limitation of our approach is that it is more suitable for professionals in a particular domain rather than for newcomers. For the last category of researchers it is more important to find main articles from a domain to get insights on what does the particular keyphrase mean, because they could mean different things such as simply co-occurring terms, sub-fields of a particular domain or different approaches, which

just occur together frequently in documents. We assume that some kind of a visualization technique could improve a novice user experience and we are planning to work on it in later releases.

Another challenge for Sci-Search system is filtering of commonly used phrases, such as “important goals”, “application requirements”, “promising applications” etc. Such filtering could be done using stop-phrases list, however now it contains not much phrases and will be updated in the future. In addition, we are planning to implement an algorithm detecting such phrases automatically and filtering them. We assume that it will increase clustering and annotating quality as well single document annotations with keyphrases.

Acknowledgments. This work is supported by the federal target program "Kadry". The program Research project 16.740.11.0751. Code of competition 2011-1.2.1-302-031. Code of application 2011-1.2.1-302-031/5.

References

1. Li, Q., Bot, R.S., Chen, X.: Incorporating Document Keyphrases. In: The 10th Americas Conference on Information Systems, New York (2004)
2. Gutwina, C., Paynterb, G., Wittenb, I., Nevill-Manningc, C., Frankb, E.: Improving browsing in digital libraries with keyphrase indexes. Journal of Decision Support Systems 27(1-2), 81–104 (1999)
3. Bernardini, A., Carpineto, C.: Full-Subtopic Retrieval with Keyphrase-Based Search Results Clustering. In: Web Intelligence and Intelligent Agent Technologies (2009)
4. Zhang, D., Dong, Y.: Semantic, hierarchical, online clustering of web search results. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) APWeb 2004. LNCS, vol. 3007, pp. 69–78. Springer, Heidelberg (2004)
5. Zeng, H.J., He, Q.C., Chen, Z., Ma, W.Y., Ma, J.: Learning to cluster web search results. In: The 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 210–217 (2004)
6. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Conference on Empirical Methods in Natural Language Processing, pp. 216–223 (2003)
7. Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. In: Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004)
8. Kim, S.N., Medelyan, O., Yen, M.: Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation, Springer Kan & Timothy Baldwin (2012)
9. Xiaojun, W., Xiao, J.: Exploiting Neighborhood Knowledge for Single Document Summarization and Keyphrase Extraction ACM Transactions on Information Systems 28(2), Article 8 (2010)
10. Zesch, T., Gurevych, I.: Approximate Matching for Evaluating Keyphrase Extraction. In: International Conference RANLP 2009, Borovets, Bulgaria, pp. 484–489 (2009)
11. Popova, S., Khodyrev, I.: Ranking in keyphrase extraction problem: is it useful to use statistics of words occurrences? Proceedings of the Kazan University Journal (2013)