Semantic Reranking of Scientific Papers According to its Research Area and Research Technologies

Tatyana Ivanova Ivanova

Abstract—Because of the rapid growth of the number of electronic publications, locating the most relevant ones to the concrete search query (and research needs) is becoming more and more difficult. The research in this paper is based on the abstract model, representing scientific research as a tuple of research area and technologies, used for performing the research in this area. The assumption is that if two scientific papers have the same research area, and use similar research technologies, they present closely related research. As main reading goals of scientists are directed in some research area, or they investigate concrete research technology, the search queries usually are intended for finding papers in narrow research area, or using specific research methods. Every of the scientific paper structural parts (as abstract, title, introduction, etc.) contain some terms, belonging to the domains of these research parts (research area and used research technologies).

In this paper we propose a vector representation of the scientific paper parts, consisting of two components: research area and research technology. These two components are numbers, represented frequency of usage of specialized terminology, related to corresponding area. We define concepts “research vector” and “relative research vector” for this purpose, analyze 90 scientific paper abstracts to determine statistically the most likely range of research vector components values, and propose scientific paper ranking algorithm, using them in reranking of scientific papers, returned from several digital libraries. We evaluate the importance of the presented paper model in query refinement stage by sending queries, classified in four groups to ACM digital library, and comparing the all returned results, and relevant ones among first 60.

Index Terms— scientific paper model, federated search, semantic search, research vector, relative research vector, ranking algorithm, metasearch engine

1 Introduction

PhD students and scientists needed from many resources in the realm of learned or studied domains: tutorials, software information and code sources, scientific papers, etc. A largest part of scientific information is stored in Digital Libraries (DL). Digital libraries are sets of electronic information resources and associated technical capabilities for creating, organizing, searching and using them by humans. Digital libraries contain information resources (papers, tutorials, etc), described by metadata, containing information about creator, owner, type of representation, reproduction, access rights, 1short domain description. Metadata also may contain links or relationships between resources and other data or metadata. Digital library resources typically are stored in databases, and hence they are deep web resources. There are valuable difficulties in crawling and indexing such resources on the one hand, and specialized searching approaches, based on specific library metadata may be used to facilitate searching in digital libraries on the other. Digital libraries typically use embedded search tools, which perform syntactic (keyword-based) search. Semantics, that user implicitly associates with the search string are not captured and used. Thus, a search query is typically broad, often ambiguous, and specific library metadata are only partially used. That is why general DL search engines return only scientific papers, containing in the abstract or title some of keywords, used in the query, and omit relevant papers, using synonyms of query words, for example. This list usually contain thousands of results, but is incomplete, as searched DL store only small part of all scientific papers, and because of natural language ambiguities, usually relevant papers appear far from the beginning of the list, after some irrelevant ones. That is why tools, that perform semantic searching of two or more DLs sending the same query, augmenting and reranking returned results is needed. As different DLs store different types metadata, and propose different ways to access it, a federated meta-search tool, making specific query refinement for every DL according to it specifics needed to be developed.

We have found several federated search engines, searching in sets of digital libraries (Infomine, Infoplease, Microsoft Academic Search, WorldWideScience, and some others), but it evaluation leads to conclusion that they return very few results in our domain (electronics, testing and diagnosis of electronic circuits), and some of them are working very slowly. That is why we decide to find the best digital libraries in our research area and create federated metasearch engine, that perform semantic searching, augmenting and reranking returned from them search results. To achieve higher precision and recall of returned results, and usage of flexible result

- *Tatyana Ivanova received his doctorate in 2009 at the Technical University of Sofia, Bulgaria, and she is an Associate Professor in the College of Energy and Electronics, Botevgrad, Bulgaria, phone: 0895589982; Bulgaria e-mail: tiv72@abv.bg*

classification strategies, we represent our domain terminology in a semantic machine-processable way, using ontologies. Presently we have developed a conceptual model of a meta-search engine for students, PHD students and scientists that uses domain and user profile ontologies, as well as information (metadata), extracted from DLs, paper titles, abstracts and web sites to improve search result quality. It is a model of a federated search engine for searching scientific resources that can make some interactive semantic query refinement, then automatically generate several search queries, sends them to previously selected digital libraries and web search engines, augments and ranks returned results, using ontologically represented domain and user metadata. For testing our model, we develop initial versions of the search engine components and domain ontologies in the electronic domain (FPGA testing and diagnosis) to represent in machine-processable way domain knowledge. For performing good augmentation and improve ranking of returned results, we extract and use domain terminology from title, abstract, and keyword sets.

In this paper we present the results from linguistic analysis of scientific paper titles and abstracts and categorization of extracted terms according to our terminological model, consisting of two components: research area and research technology. We use this categorization to represent paper abstract and title terminology in the way, closely related to the scientists view. We analyze scientific paper abstracts to determine the relative part of the every type terminology, and propose scientific paper ranking algorithm, using proposed model and scientific terminology specifics.

2 Related Work

Adding semantics in searching technologies is useful both for query formulation (interactive or automatic refinement, user assistance), and ranking of returned results. There are two main approaches for search engines to present results: Ranking them in one list, on the base of it popularity or relevance to the query (or other criteria, as issue data, citation numbers, etc.), or firstly clustering them in several groups, then ranking results independently in every group, and show (hierarchical structure or list of) these groups.

Yippy (http://yippy.com/) is good web metasearch engine, clustering returned results on the base of it snippets, proposing by search engines, and presenting results organized in sets of thematic clusters. Some scientific metasearch engines as WorldWideScience.org or Microsoft Academic Search also propose clustered results (not only thematic, but by authors, publication date, place, etc.).

Scientific papers have several specific features that can be used in the search: relatively fixed structure (as textual documents, having title, abstract, subtitles, abstract, and conclusion), thematically well-formulated titles, keywords, abstracts, bibliographical information, citations, and well defined domain vocabulary (restricted natural language is used). There are a few researches on searching scientific papers, using these features.

TheWEBFIND approach [8] uses reliable external sources (MELVYL and NETFIND) to yield bibliographic records and paper author's information from the web.

FutureRank[2] uses the authorship and citation network, and the publication time of the article in order to predict future citations. FutureRank is accurate and useful for indexing and ranking publications on the base of well – predicted citations (instead of widely – used real, but old data on citations).

Many information retrieval algorithms, as cluster-based retrieval [7], Relevance Feedbacks based, graph analysis algorithms [3] such as Page Rank [5] and HITS [4], have been used for scientific document search, ranking returned results, based on user query and, question answering systems [6].

DT-Tree (DocumentTerm-Tree) [1] is a effective clustering algorithm, based on scientific paper structure, and used relatively – small dimensions when representing document as a vector

Arnetminer (arnetminer.org) aims to provide comprehensive search and mining services for researcher social networks. It can determine the relevant topics or subtopics to the user query create a semantic-based profile for each researcher by extracting information from the distributed Web, search heterogeneous scientist's social networks, integrate academic data (bibliographic data and researcher profiles) from multiple sources. It can provide scientific paper ranking information (related to authors, citations, impact factor, etc) and rank papers, using 8 ranking metrics.

Some of ontology matching techniques are based on calculation of concept similarity, using taxonomic structure or non-taxonomic relations. Such class of concept similarity measures uses ontological representation of the domain as labeled graphs and compares nodes in these graphs using lexical and structural features [12].

Using semantic representation of domain knowledge in the ranking process is successfully experimented and applied in medical and biological domains. The system GoPubMed [5], uses three ontologies - the Gene Ontology, MeSH and Uniprot, to improve PubMed searching query results. Results are presented in a faceted hierarchy that includes ontology terms, authors, journals, and publication dates. The system allows the user to navigate by each of MeSH terms. Unlike in our work, no attempt is made to reconcile the contributions of multiple terms into a single score.

In [13] the research is focused on adding the semantic dimension to biologic and medical literature search, in PubMed, the most significant bibliographic source in life sciences This work explores ways to use high-quality semantic annotations on the base of MeSH vocabulary to rank returned search results. Several ranking functions that relate a search query to a document's annotationshave been developed and an efficient adaptive ranking mechanism for each of them is proposed and tested. It also describes a two-dimensional Skyline-based visualization that can be used in conjunction with the ranking to further improve the user's interaction with the system, and demonstrate how such Skylines can be computed adaptively and efficiently.

Using of semantic technologies is approve it potential for significant improvement of search results in some search engines (for example google, ask) and in some domain DLs (for example medical and biological), but we have been found no one semantic content-based search tool for scientists in the domain of electronics.

Different above discussed approaches make attention on different scientific paper properties (quality related, as citations; structural and bibliographic; and semantic, as research domain semantic models-based ones). As all these researches have it scientific achievements, we believe, that all that paper characteristics are important in the searching and ranking process. Our main idea is to use all above discussed paper characteristics (structural, quality-based and semantic) in developing of searching tool for scientific papers. As semantic approaches are to some extent doain-dependent, we will take in account semantic searching approaches, proven in other domains, and semantic ontological representation of our domain terminology in our work.

3 Scientific Paper Title And Abstract Terminology Analycis

In this chapter we will present a terminological model of scientific paper, closely related to the real semantic paper structure (that scientist have in mind, when writing it), check the correctness of this model by making terminological analysis of title and abstract of many scientific papers from digital libraries, and make preliminary estimation of possibilities of it usage in the process of ranking or clustering returned from search query results.

Scientific digital libraries store much bibliographical and other type information about stores papers, but there are different license agreements for usage of this information for different libraries. In our work we will use only information, explicitly shown in search result list, returned from native DL search engine. There are significant differences in information, related to returned results for different libraries. In ACM, for example, keywords are not used at all. Instead of them, the index terms (that are elements of the general ACM classification schema) are used, but many papers have no index terms. ACM proposes one four-level Primary Classification, and one, Additional Classification which may contain 1,2,3, or more upper-level indexes, every of them having one or more sub indexes. This provides many terminological information for indexed papers, contrary to the papers, having no indexes, or IEEE – returned results, while keywords are missing at all. That is who in our research we will use only title and abstract terminology to ensure common basis for terminological analysis of results, returned from different scientific DLs.

3.1 Scientific paper title and abstract terminology characteristics

The title should be indicate or describe the contents of the article, and hence, usually contains valuable domain terms or phrases. The abstract is a summary of the work, and is intended to serve as a guide for the article purpose, content, achievements, and to furnish subject metadata for indexing services. That is who it should contain high percentage domain terminology. Usage of abbreviations and acronyms is typical for scientific abstracts. Some abbreviations are standard (for example hr, min, sec, etc) and are used without definition. Using abbreviations is recommended in scientific text, and non standard ones should be defined in it first usage. Sometimes definitions are omitted (especially for frequently-used or popular abbreviations). Our first aim is to evaluate the expected percentage of domain terminology in scientific paper abstracts. As closely related domains usually contain some common concepts, conclusion about the paper domain may be wrong, when it is based on the usage of very few domain terms. What percentage of domain terms will guarantee a high precision of the conclusion about paper domain?

We first will divide paper abstract and title terms in two main classes: function words and content words. Then we divide a class of content words in two groups: common research words (CRW), that are used in scientific papers to describe scientific research, and domain terms. Function words are words that have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with other words within a sentence, organize grammatical relationships between words within a sentence, or specify the attitude or mood of the speaker. They signal the structural relationships that words have to one another and are the glue that holds sentences together. They serve as important elements to the structures of sentences. There are a relatively small and fixed number of function words. These are prepositions, conjunctions, determiners, pronouns, and auxiliary verbs. We download a comprehensive list of these words from http://www.sequencepublishing.com/cgi-bin/download.cgi?efw and use it to remove function words from paper titles and abstracts before analyzing it domain terminology.

Common scientific words and expressions (also known as Academic Words) are domain independent, opposite to specific domain terms that denote concepts, objects, and processes of the particular scientific domain. They are used to design and organize scientific text by connecting text fragments devoted to different topics and subtopics or by expressing the logic of reasoning. We download from http://www.sequencepublishing.com/academic.html and use the Academic Word List (AWL) for extracting research domain independent scientific words. It contains 570 words, specific to academic texts where they account for about 9-10% of running words, divided into several categories and sorted by frequency of usage.

There are several scientific paper types [9]: Original articles; Case reports; Technical note; Pictorial essay; Review;

Commentary; Editorial; Letter to the editor; and some others. We will analyze first in detail the abstract terminology of the two types: Review, and original articles, as some of the others are similar to some of them in terminological point of view, or are not so important as a scientific paper.

Original Article is the most important type of paper. It provides new research results based on performed research. It describe two main components of scientific research: Results (saying what was found and where it can be used), and Methods (saying what technology or instruments are used to achieve the results). In this type of papers several domain (usually closely-related) terminology are used. For example, in the paper, described ontology-based e-learning method terms from e-learning, ontological knowledge representation, and some general computer science-related domain are used. Related to the main paper topics for students or scientists may be other ontology-based methods (not intended for e-learning), or other e-learning methods (not only ontology-based) and easy reachability of these classes of scientific papers will be useful in many cases. This will be easy, if we distinguish the research methods or tools, and research domain terminology from each-other. In the above example, research domain terminology is e-learning terminology, and research methods is a semantic web, or ontology terminology, and them may clearly distinguish, but in some cases doing this is not so easy.

To take the experimental view on the research paper abstract terminology classification, and make initial experimental verification of our scientific paper terminological model we make following experiment: using google scholar, we collect 30 paper abstracts, discussing researches, related to ontologies in e-learning, 30 paper abstracts, discussing researches, related to ontology usage in biology, and 30 paper abstracts, discussing researches, related to ontologies (mapping, learning, management and so on). We carefully read all these abstracts and make manual classification of it terminology, according to our model. Average results (in percents and different term average count) are shown in table 1 and graphically presented in fig.1.

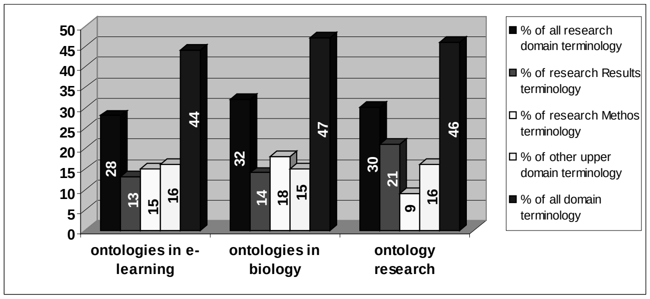

Fig. 1. Scientific papers terminology classification – graphical representation

Table 1 – Scientific papers terminology classification according to research area and technologies model

	% of research results terminology	% of research methods terminology	% of all research domain terminology	% of other upper domain terminology	% of all domain terminology
ontologies in e-learning	13	15	28	16	44
ontologies in biology	14	18	32	15	47
ontology research	21	9	30	16	46

Main conclusions of our experiment are:

1. Average of all domain-related terms in scientific paper abstracts is at about 45-48 %. This corresponding well to the other scientific paper terminology research, showing that at about 10% of it terminology are common research terms, and common words are 35-40 %.
2. Average of all research domain terms in scientific paper abstracts is at about 30 -35%.
3. Research domain terms typically are belonging to 2 (may be closely related) domains: research results domain, in which results are described, discussed, or used, and research methods domain, describing technologies, algorithms, or tools, by mans of which results are obtained. The ratio of the number of terms in these two areas is variable but, their total number is at about 30 - 35 % of the all words in the abstracts.

3.2 Paper abstract research vector and paper title research vector, and relative research vectors

Based on these experimental results, we will define vector representation of the scientific paper abstract, consisting of two elements: paper research area (results domain) and paper research technology (methods domain), and give it name “paper abstract research vector”, or simply “research vector”. So,

Definition 1: a tuple (percent of research area terminology, percent of research methods terminology), calculated for scientific paper abstract we will say “relative paper abstract research vector”, or simply “relative research vector” and denote as RPA (RAA, RMA).

Analogical experiment for paper titles leads to the conclusion, that this term will be correct also for paper titles. So:

Definition 2: a tuple (percent of research area terminology, percent of research methods terminology), calculated for scientific paper title we will say “relative paper title research vector”, and denote as RPTV (RAT, RMT).

In some abstracts, a few (in some cases one or two) domain words are used many times. In such cases, percent of domain terminology may be significant resulting from repetition, but small number of terms actually used is a prerequisite for errors in determining the domain. That is why the number of used different terms also is important, and we will define also research vectors:

Definition 3: a tuple (number of different research area terms, number of different methods terminology), calculated for scientific paper abstract we will say "paper abstract research vector", or simply "research vector" and denote as PAV (AA, MA).

Definition 4: a tuple (number of different research Results terms, number of different Methods terminology), calculated for scientific paper title we will say "paper title research vector", and denote as PTV (AT, MT).

While relative research vectors show the percentage of research – related terminology in the paper parts, Research vectors contain the number of used different terms, and may be used to determine whether extracted from the paper parts domain related terms are sufficient to identify it research domain. For example, if one domain term is used 7-8 times in short abstract, relative vector for the domain may have good value, but if this term is used in some other domains, this is not sufficient for domain identification.

These terms will be correct also for paper keywords, but it is out of reach of our current research.

As it is clear, if one term is used 2 or more times in the abstract, corresponding relative research vector component is as great, as many times it is used, but corresponding research vector component is independent from it times of usage. When relative research vector component is relatively big, but research vector component is 1, or 2, this means that 1 or 2 domain terms are used many times in the abstract, and information for clear recognition of the underlined domain in the abstract is possibly insufficient.

4 The Clustering And Ranking Algorithm

On the base of cited above linguistic researches and our experimental results, we present the following terminological model of scientific paper's parts (title or abstract) (Fig.2). As above experiment shows, in every paper part there are domain independent terms and domain dependent terms, in approximately equal proportions. Domain independent terms include common words and common research words. Domain dependent terminology includes upper domain terminology and Research vector, containing paper research area word and paper research technology words. We will use the quantitative characteristics of this model in the process of ranking and clustering the returned results, as well as in some query refinement techniques.

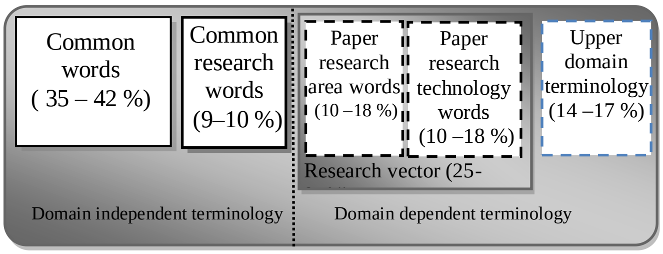
Common words (35 – 42 %)
Common research words (9–10 %)
Paper research area words (10 –18 %)
Paper research technology words (10 –18 %)
Upper domain terminology (14 –17 %)
Research vector (25-
Domain independent terminology
Domain dependent terminology

Figure 2. The terminological model of scientific paper's parts

This model has three main advantages:

1. It clearly describes scientific paper part terminology as augmentation of different type words
2. Research vector is terminology representation very near to the researcher view on the paper
3. It clearly show types of words, recommended for usage in a good search query (common and upper domain words must not used, and for good query formulation, at least one unambiguous term for each research vector component should be used

The main disadvantage of the model is the fact, that average percentages of the tree types of domain – related terminology may vary significantly or unpredictable from paper to paper.

Another potential disadvantage is the need of previously developed knowledge-based resources. Upper domain terminology should be organized in thesaurus, or ontology. For it development freely available resources in the internet (as general dictionaries, thesauruses, or hierarchical classifications) may be used. Ontological representation is recommended, as it will guarantee higher precision in abstract and title terminology analysis.

Domain ontologies should be carefully developed, and interpretations of paper domain terminology as Paper research area or Paper research technology should be done dynamically, as a temporal mapping between user profile and domain ontology, inspired from interactive user query refinement. Domain ontologies should represent domain terminology relationships, as hierarchical relations, other semantic relations, synonymy. As a whole, in every scientific research domain research objects and research methods are clearly defined and separated, and this separation should be explicitly presented in the domain ontology. For example, devices and testing methods in our ontology are hierarchically independent (they are in different branches of the ontology hierarchy tree), and semantic relations between them are implemented, showing possibilities that the device may be tested, using selected method.

Proposed in this paper searching and ranking approach is based on above presented paper parts terminological model, and it main ideas are to:

1. Propose adequate guidance to the scientist to help him in writing search queries, representing the research vector of it scientific needs
2. Propose syntactic and semantic linguistic analysis tools for precise terminology extraction from paper titles and abstracts, removing or domain-independent terms, and in such a way, constructing the set of domain – related terms of every paper title or abstract
3. Propose semantic tool for determining if every such term is belonging to the presented in the search query research area or research technology domain. All belonging terms will be used to calculate possible research vector coordinates.
4. If the two parts of calculated relative research vector are in the intervals, shown in the model, paper is classified as appropriate for the query and will be ranked in the cluster of relevant papers
5. If only one part of calculated relative research vector is in the interval, shown in the model, paper is classified as partially appropriate for the query and will be first clustered in the set of all results, having only this research vector component in range, and then ranked in it determined in such a way cluster.
6. Results will shown in four main clusters:
   - Cluster of sorted relevant results
   - Cluster of partial relevant on the base of the right research area results, having internal clustering according to various research technology types (one subcluster for each research technology type)
   - Cluster of partial relevant on the base of the right research technology results, having internal clustering according to various research areas (one subcluster for each research area).
   - Cluster of all other returned results which also may have sub clusters

Ranking assumptions:

1. Relevant term in the title is more important than in abstract
2. Citation is important, as it is the criteria for paper quality, but it is not more important than domain relevance
3. Following citation assumptions are important:
   - More cited papers are more useful.
   - Mutual reinforcement between articles and the authors is obvious, but calculations take more time, and we will not take account on this.
   - Recently published articles are more useful, or in other words, they will obtain more citations in the future.

Having in mind these assumptions, we propose the following ranking formula:

RP = RAA + RMA + α*(RAT + RMT)+ CIT* CITN,

where:

RAA, RMA, RAT, RMT, are explained in definition 1 and definition 2;

α is a coefficient; α should be greater than 1 (α>1), as it presents relative importance of title terms according to the abstract terms.

CIT is the citation importance of the paper. To calculate it, we use the sequence of coefficients between 0 and 1, calculated using the dependence between year of publishing and number of citations;

CITN is number of all paper citations.

Calculation of, RAA, RMA, RAT, RMT is made on the base of vector space model in information retrieval [14], [15] including the following modifications:

1. initial weight of every meeting of the term is 1
2. for every pair of terms we increase by 1 it sum of weights, if they are associated by some relation in the domain ontology, and are in different sentences in the text
3. for every pair of terms we increase by 2 it sum of weights, if they are associated by some relation in the domain ontology, and are in one sentence in the text.

In such a way, we enrich traditional vector space model by adding semantics, using previously developed domain ontology.

So, the proposed ranking function include semantic, or content-based elements, structural ones (as the weight of the terms depends of it inclusion in the title or abstract), and link-based elements (expressed by citation part CIT* CITN).

Advantages of using such aggregated ranking function are that it includes information of almost all characteristics, valuable for searching and ranking. For example, if two papers have almost equal domain relevance, the more cited one will appear before other in the list, or if two papers have almost equal number of citations, the more closed to the research domain will appear before other in the list.

Disadvantages of using such aggregated ranking function are that papers, having many citations will appear in the beginning of the list, because of the grand value of the citation part CIT* CITN, even if they are not relevant to the research domain. The mechanism of calculating the CIT may reduce these cases, but other approaches to minimize these possible errors, as including additional coefficient, or threshold of the value of this component should be evaluated.

Another problem is that this generalized function hide relative amount of the research area and research methods components. The concrete values of the research area and research methods components are important mostly in the clustering process. In this case, metrics of type:

RPA = RAA + α* RAT + CIT* CITN for ranking in the research area clusters

And

RPM = RMA + α* RMT+ CIT* CITN for ranking in the research methods clusters should be used.

We evaluate the importance of the above model in query refinement stage by sending to ACM 20 groups of queries, each of which is represented in 4 variants – by using terms for two components of the research vector, and upper domain term, using terms for two components of the research vector, or only for each component. I test and evaluate 32 of all queries, describing clearly and explicitly my intent in terms of the research area and research technology, sending queries and after brief reading of the returned results qualifying them as relevant or irrelevant. For another part of results I gave queries to the students, ending course of testing and diagnosis of integral circuits. I gave them these queries, and comprehensive description of the searching goals, and ask for evaluating the returned resource relevance.

Two examples and average results are shown in table 2, and average results are represented graphically on Fig. 3 and Fig. 4. Conclusions are that query refinement by upper domain term (when it include keywords, presenting both components of the research vector) is not recommended as a whole, as it does not improve the quality of results (in some cases, in first60 there are more relevant results, but these are few queries, and usage of some upper domain terms lead to significant deterioration of the results). Refinement by using hypernyms not everything lead to better results even in case of short and ambiguous queries, containing keywords, related to only one of the research vector components.

User feedback may be used to store good for query refinement hypernyms. Using at least one term belonging to each component of research vector is strongly recommended, as it leads to better filtering and ranking of returned results. As average, returned results are significantly less, but in many cases sufficient, and are arranged much better. Synonyms in disjunctive queries should be used to increase the recall of refined query. In case of using only one vector component, there usually is a grand amount of irrelevant results, precision is very low. So, the presented model of representation of scientific paper in terms of research area and research method is in the base of useful query refinement strategy.

Table 2. Query types – examples and summary

Query /type	All returned results	Relevant among first 60
Example 1		
e-learning system interoperability	1115	55
e-learning interoperability	1136	59
e-learning	11456	17
interoperability	24750	2
Example 2		
ontology mapping e-learning system	569	54
ontology mapping e-learning	572	57
e-learning	11456	17
ontology mapping	10881	3
Average from 20 examples		
including upper term	975	51
only 2 vector component	963	57
only first component	9385	13
only second component	13875	9

We have made partial evaluation of our reranking algorithm, based on upper defined ranking metric. Our experiments are made by using two-component vector queries, sending to ACM. We have sent 50 queries and manually check result of reranking in first 60 results. Our first goal is to find the best value for α. We try 1, 1.5, 2, 2.5, 3, and 3.5. For α = 2 average reranking effect was the best.

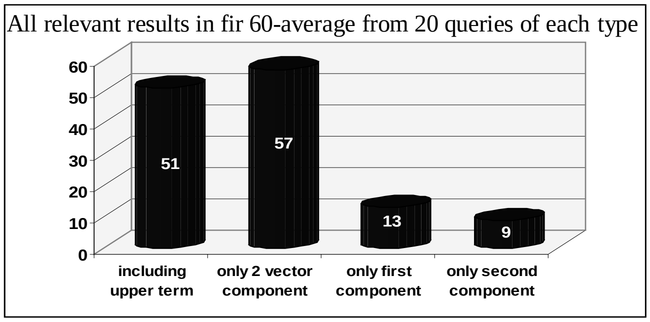

Figure 3. Average of all relevant among first 60 results of four query types

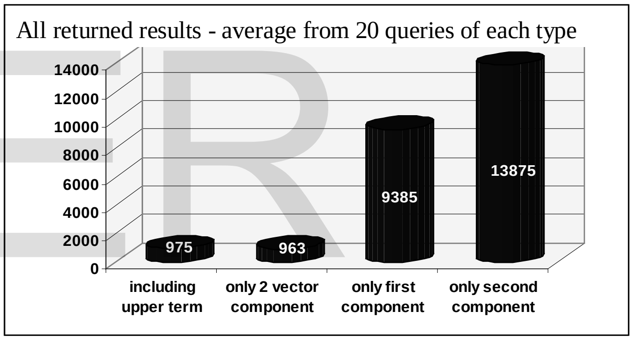

Figure 4. Average of all returned results of four query types

The most important strength of the proposed model is that it can support scientific paper clustering according to it research area and research technology. We are working on implementing the following scenario:

In the stage of query formulation, query refinement tool supports explicit refinement of research area and research method query components, using Ajax communication methods and domain ontology. Possible semantically relate terms are proposed to the user and he/she choose the best ones for expressing his search intent. Parallel to the sending of the query, the system extracts the context of it terms from the domain ontology for future comparison to the extracted from returned results titles and abstracts.

The first ten returned results from the search engine are shown to the user to avoid performance problems, and in this time all the result abstracts and titles are processing by domain terminology extraction module and research area and research method vectors are calculated, using extracted contexts. Terms, belonging to the contexts, and extracted from the chosen result are considered as part of the corresponding vector. For papers, having percentage of research area or research methods terminology under 10%, extracted domain terminology is searched in domain ontology for establishing of another area or method, or concluding, that it research area or method is outside of the interested domain.

After calculating RPA and RPM metrics for every paper and it research area and methods, papers, having our research area a sorted according to RPA, it research methods and RPMs are displayed. Sorting according to RPM also can be made, but we think it is of less importance.

Such representation of results is very useful, because it shows not only relevant papers to the query, but also closely related researches areas and methods, and papers, presenting them. It is useful both for future making of clear and effective searching queries for scientific publications, and for supporting extending of the scientists research area.

5 Conclusion

In this paper we propose a vector representation model of the scientific paper parts, consisting of two elements: paper research area and paper research technology. We define concepts "research vector" and "relative research vector" for this purpose, analyze scientific paper abstracts, determine experimentally the most likely range of research vector components values, and propose scientific paper ranking and clustering approach, using them in reranking of papers, returned from several digital libraries. We define the reranking metric for the ranking algorithm and perform some partial evaluation of this metric. The drawbacks of our metrics are two: it doesn't take in account relations between terms, and its multiple usages. This may lead to mistakes in cases, when used terms are not related, and a few terms are used many times in the abstract for example. We will test these potential variants and improve our metric, if needed.

When the value of one of components of the relative research vector is small, it is very likely that it corresponding domain is different from the searching. For such cases, we will make research to find appropriate clustering algorithm to group corresponding results according to its domains.

The main drawback of reranking is it low speed, as it relies on terminology extraction from grand number of abstracts or titles. This is the well-known problem of many metasearch engines. We will experiment strategies for partial reranking or clustering to improve the performance.

References

[1] S. R. Ali Rizvi , S. X. Wang, "DT-Tree: A Semantic Representation of Scientific Papers", 10th IEEE International Conference on Computer and Information Technology, 2010
[2] H. Sayyadi and L. Getoor. FutureRank: Ranking Scientific Articles by Predicting their Future PageRank. In Proceedings of SDM'. pp.533~544 , 2009.
[3] J. Lin., "PageRank without Hyperlinks: Reranking with PubMed Related Article Networks for Biomedical Text Retrieval." BMC Bioinformatics, 9:270, 2008.
[4] JM. Kleinberg, "Authoritative Sources in a Hyperlinked Environment." Journal of the ACM, 46(5) pp. 604-632, 1999
[5] O. Kurland, L. Lee , "PageRank without Hyperlinks: Structural Re-Ranking using Links Induced by Language Models." , Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), Salvador, pp.306-313, 2005.
[6] J. Lin and D. Demner-Fushman, "Semantic Clustering of Answers to Clinical Questions." Proceedings of the 2007 Annual Symposium of the American Medical Informatics Association (AMIA 2007), pp. 458-462, 2007.
[7] X. Liu, WB. Croft, "Cluster-Based Retrieval Using Language Models." Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), pp. 186-193, 2004
[8] A. E. Monge, and C.P. Elkan, "The WEBFIND tool for finding scientific papers over the worldwide web", Proceedings of the 3rd International Congress on Computer Science Research., 1996
[9] Peh WCG, Ng KH., "Basic structures and types of scientific papers", Singapore Med J., 2008.
[10] H. Sayyadi, L. Getoory, "FutureRank: Ranking Scientific Articles by Predicting their Future PageRank", In Proc. of the 9th SIAM International Conference on Data Mining, 2009.
[11] A. Doms and M. Schroeder. "GoPubMed: Exploring PubMed with the GeneOntology". 33, 2005.
[12] J. David and J. Euzenat. "Comparison between ontology distances (preliminary results)". In International Semantic Web Conference, 2008.
[13] J. Stoyanovich, W. Mee, and K. A. Ross. "Semantic ranking and result visualization for life sciences publications". Columbia University Technical Report cucs-028-09, 2009.
[14] http://en.wikipedia.org/wiki/Vector_space_model
[15] J. Datta, Ranking in Information Retrieval, MTech Seminar Report, www.cse.iitb.ac.in/.../TR-CSE-2010-31.pdf , 2010