Semantic Search for Scientific Publications Based on Rhetorical Structure

Lan Huang, Kai Feng, and Hao Xu*

College of Computer Science and Technology, Jilin University, Qianjin Street 2699, Changchun, China
huanglan@jlu.edu.cn, fengkai15@mails.jlu.edu.cn, xuhao@jlu.edu.cn
*Corresponding Author

Abstract. Most scientific papers have their own rhetorical structures, which have deeply rooted in the minds of both authors and readers, such as background, problem and discussion. However, most existing search engines for scientific publications haven't made good use of such semantic information. In fact, each reader would be interested in different semantic modules of a paper, that is, certain concepts or entities mentioned in different semantic parts represent various indications. In this paper, we design and implement a semantic search platform that aims to provide semantic search for scientific publications based on rhetorical structure. To provide better results, we initiate with the semantic model of scientific papers, so as to meet the special attention of the semantic module in papers for readers.

Keywords: semantic search, rhetorical structure, semantic annotation.

1 Introduction

Authors always hold an logical structures in their minds while they write scientific papers. Besides, each scientific paper has its own rhetorical structure such as research background, problem statement, solution and future work. For the sake of readers to read more targeted, some publishers even require the structured abstract that authors must provide. Nevertheless, the traditional search does not have strong connections with these semantic information and numerous metadata of articles. Actually, according to the phrases or words attained from the users', the search engine positions the documents. These words are regarded as the ordinary characters rather than any concepts.

Semantic search is an application of Semantic Web [1]. Naturally, semantic search for scientific papers needs a huge database for publications and metadata. The model of data becomes the fundamental point in achieving semantic search because of the structured data, which could be read by machines, and it is also convenient for us to compute the relations between them [5]. Several models have been devised to label the rhetorical structure within the papers. The Harmsze Model proposes that a paper is constituted by metadata, positioning, methods, results, interpretation and outcome [2]. Another model for science publications named ABCDE includes annotations, background, contribution, discussion and entities [6]. Thus, most papers could be labeled

by these two coarse grained models. Generally, a paper contains modules, which are made up by metadata, background, problem, solution and discussion. And the main problem is the detection of rhetorical components. With structured data and relationships stored in the database, semantic search services could be provided. ClaimFinder was a research prototype which delivered the search services based on the original data [3]. The home page of this website allows users to do the keyword search and shows the result about the concept and some relations linked in the concept. And Mimir, an Open-Source Semantic Search Framework, could provide complex queries on account of natural language process and this framework is built on a cloud storage platform [4]. It stores annotations, tokens, index of all the basic data and etc. Thus, it would provide the better result than the traditional way.

For scientific publications, the meaning of a concept would be vary when it appears in different semantic modules. It is worth mentioning that different people may pay attention to the different parts of papers. In view of that, we design and implement a semantic search platform based on the rhetorical structure and natural language processing and semantic technologies. The platform extracts the keywords in different semantic modules of papers. Meanwhile, semantic search could use these keywords and readers could choose the semantic module they prefer. The platform would do the search under the rhetorical structure and retrieve the list of papers which are more accurate. Our goal is to provide more efficient and effective search services.

2 System Design

In order to accomplish the task as much as possible, we divide the work into three parts: semantic annotation, concept/entity detection and semantic search. Semantic annotation is the work for adding labels of rhetorical structure for scientific papers. Concept and entity detection is to extract the keywords under rhetorical structure and to store them. The part of semantic search is to process keywords and to compute the results based on semantic modules.

The structure of the system as the Figure 1 shows below.

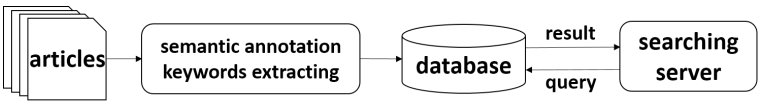

Fig. 1. Overall architecture of the system.

The data of articles contains the title, authors, year, and the whole text of the article and they need to be processed by means of the semantic annotation and concept and entity detection. Then, the structured data of each article ,including the keywords of every semantic module and the title of the article, would be stored in the database via the system. Finally,the database would be able to accept the requirements through searching server and to return the result.

In this passage, the system contains two main functions. The first one is searching papers based on rhetorical structure. The second one is enabling readers to annotate

semantic modules for scientific papers at their wills. The system would take the measure of statistics to achieve the final semantic modules.

2.1 Semantic annotation

Before semantic annotation, we patternized basic rhetorical structure of scientific publications that could be modified or extended. The experiment of this passage pays emphasis on general rhetorical structure of scientific papers. For the convenience of us to discuss, we entitled them as background, problem and solution respectively and there are two ways to make semantic annotation. The first way needs the help of readers. As the Figure 2 shows below, once the button clicked, the webpage would copy the text selected and send it to the server. Then the system would extract the keywords through the information the system stored.

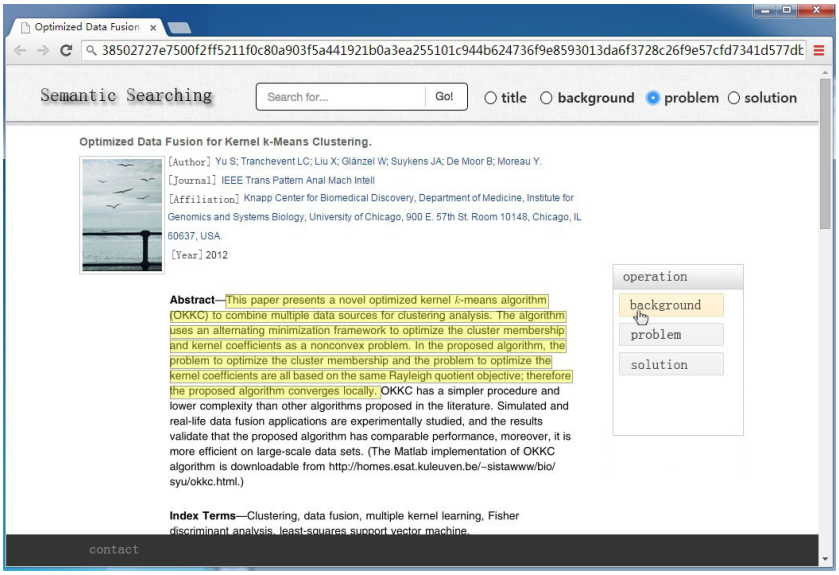

Fig. 2. Readers complete the semantic annotation.

The second one to make the semantic annotation is achieved on the basis of LaTeX, which could manage text with labels. Meanwhile, we have designed some labels like "\background", "\problem" and "\solution". If the authors use them, the system could extract the keywords through the labels.

2.2 Concept and entity detection

The same concept could be distinguished easily for its different meaning in different rhetorical structure. The system should guarantee that the concept and entity detection relies on the accurate rhetorical structures.

In order to extract keywords, the system uses a program named Jieba word segmentation, which is a java program for word segmentation. The traditional technique the program used is named TF-IDF, which made good use of simple but effective ways to extract the keywords. According to the term frequency and inverse document frequency of the words appearing in the document, the system filter out common words and preserve vital words. Once the system got the keywords, it would store them with the link pointing to the article in the database as the Figure 3 shows.

	_id	back_key	prob_key	solu_key	title
1	0	Semanti...	machine...	complex...	Learning...
2	1	XML,NLP	XML,NL...	annotati...	XML-Ba...
3	2	linguisti...	NLP,co...	NLP,gap	Matchin...
4	3	querying...	ranking ...	ranks,rel...	SemRan...
5	4	semanti...	semanti...	complex...	iMAP: Di...
6	5	distribut...	Hadoop,...	high per...	Data Mi...
7	6	data min...	incorrec...	bioinfor...	Towards...
8	7	data min...	k-means...	k-means...	A PTAS ...
9	8	data min...	practical...	smoothe...	Smooth...

Fig. 3. The attributes to store the keywords.

The first column is an identifier for each article. The second column is about the keywords of ”background module”. The next column deal with ”problem module” and the fourth column concerns ”solution module”. The fifth column stores the title of article. It presents the results in descending order by the score that ranks the statistical significance in each semantic module. It would have fifteen to twenty keywords in each semantic module of each article.

3 Semantic Search

Semantic search provides insight into unstructured documents stored by extracting the relevant keywords and index statistics in the database. Then, it is also used to identify these keywords and index similar or related documents.

According to the keywords extracted before, the platform could retrieve the keywords in the semantic modules. In this passage, we take the score into full consideration by both the weight and the published year. Since the keywords of a semantic module of each article at the level of importance is not the same. To be specific, the hit keyword in the first position of the article is assigned to three points. Then, the second one represents two points and the following keywords are all amounted one point. Besides, the score of published year is two in the past decade. And the articles published ten years before amounted one point. Through the calculation of the searching server, it

then return the list of articles. The communication of data between the readers and the database is as Figure 4 shows below.

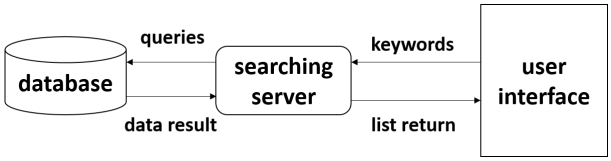

Fig. 4. The process of query requirement.

First, readers input the keywords. The searching server sends a query requirement to database. Then the database returns the data to searching server. Finally, through the processing of the server, an ordered list would return to readers.

Readers could search the keyword and decide which semantic module they prefer. So when the platform shows the result, it could focus on the scientific papers to hit the keywords in the semantic module selected just now. For instance, when readers search the keywords under the "solution module", semantic search would search the keywords in the solution column (key_solution) of database and return all the papers hit the keywords in their solutions. As the Figure 5 shows, the platform lists all the papers using clustering to solve some problems.

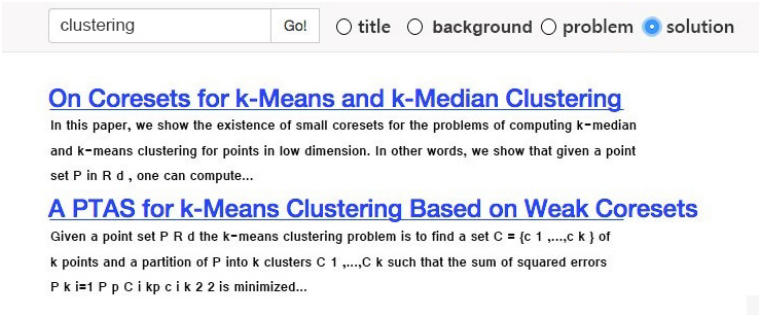

Fig. 5. The result semantic search returned.

Besides, the platform could search in another way. With the help of semantic matching, the platform could calculate all the relationships among the keywords, including "more general" and "less general" relationship. In this experiment, when the platform is searching "data mining", an auxiliary program for semantic matching would find all the words which have relationships with "data mining". And Figure 6 shows the result when readers searched "data mining" as the keywords. Since "k-means" is less general than "data mining", some papers about k-means will be returned.

4 Conclusion and Future Work

The traditional search returns the correct results and also returns too many of the articles that are not so accurate. And the semantic search narrowed the range of the paper listed

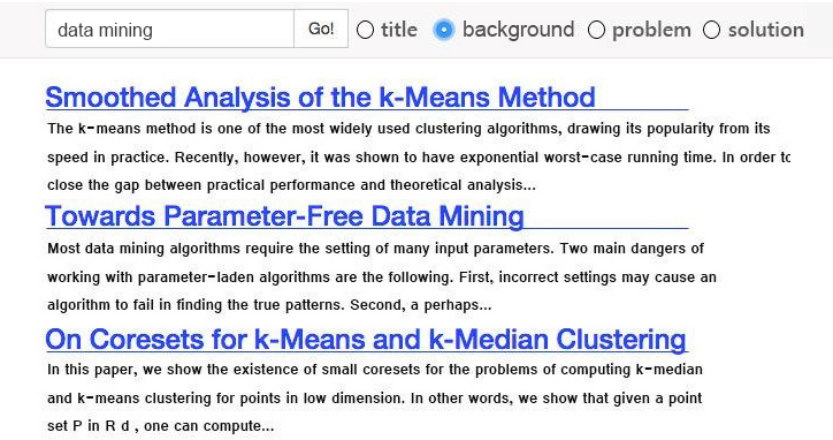

Fig. 6. The semantic search result.

through the semantic annotation. From this view, searching scientific articles based on the rhetorical structure becomes more rapid and accurate.

The platform is a preliminary experiment. And the next phase of work is to label entity and implement the across-language platform. By tagging entity, readers could understand the involved concept easily and find articles more accurate.

5 Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61300147), China Postdoctoral Science Foundation (No. 2014M551185), and Science and Technology Program of Changchun (No. 14GH014).

References

1. Guha, R., Mccool, R., Miller, E.: Semantic search. International World Wide Web Conference 36(1), 700–709 (2003)
2. Harmsze, F.A.P.: A modular structure for scientific articles in an electronic environment. Ph.D. thesis, University of Amsterdam (2000)
3. Shum, S.B., Clark, T., Groza, T., Handschuh, S., 09gnes Sndor: Scientific discourse on the semantic web: A survey of models and enabling technologies. Semantic Web Journal Interoperability Usability Applicability (2010)
4. Tablan, V., Bontcheva, K., Roberts, I., Cunningham, H.: Mimir: An open-source semantic search framework for interactive information seeking and discovery. Web Semantics Science Services & Agents on the World Wide Web p. 52C68 (2015)
5. Tran, T., Cimiano, P., Rudolph, S., Studer, R.: Ontology-based interpretation of keywords for semantic search. Lecture Notes in Computer Science 4825, 523 (2007)
6. de Waard, A., Tel, G.: The abcde format enabling semantic conference proceedings. In: SemWiki (2006)