

Literature Search Framework by Analyzing Key Aspects

Zhuyue Si

School of Computer Science
Beijing Institute of Technology
Beijing, China
Email: sizhuyue@bit.edu.cn

Gaokun Dai

School of Computer Science
Beijing Institute of Technology
Beijing, China
Email: dgkmao2340465@126.com

Zhendong Niu

School of Computer Science
Beijing Institute of Technology
Beijing, China
Email: zniu@bit.edu.cn

Abstract—Quickly moving to an unfamiliar field for researchers is painful due to a mass of scientific articles to study without knowing their research contents. Research point and its implementation method are two helpful aspects. However these two aspects are difficult to be achieved without appropriate semantic tools. This paper proposes a literature search framework which supports a new perspective retrieval of research point and implementation method of articles. This framework contains four components: extracting the above two aspects from articles, analyzing them by semantic similarity, ontology population and index building. The experimental results verify the effectiveness of this framework. And we apply this framework to the field of computer science, demonstrate the prototype system.

Keywords—literature search; information extraction; semantic similarity; framework

I. INTRODUCTION

When researchers come into an unfamiliar field, what they are most concerned about is recognizing the research contents of articles easily. Research points or method used in an article could help a lot for fast learning. However these two aspects are difficult to be analyzed and accessed without appropriate technological tools.

Describing articles using the above key aspects in semantic level needs to establish rich insight of scientific research. It requires an understanding of sentence and argument structure which is a work of information extraction, and it's more than an understanding of the "topics"[3] of articles, which was used in the previous work, such as studying the scholarly impact[7] and the history of ideas[2] of articles. There are some prior works on extracting key aspects from articles, Yuka and Yo[16] use "what can this technology be used for", "how good is the system for such-and-such type of data?" as the role of entities and events that are described in academic literature, and design an annotation scheme for formalizing relation structures in research papers. Fukuda et al.[6] extract the TECHNOLOGY and EFFECT from patent and research papers.

Only extracting key aspects from articles is not enough, semantic analysis is also essential, and this is a lack in the above works. Semantic web technologies[14], particularly ontology-based approaches, could be used for the comprehensive meta-level modeling and retrieval. Amini, Ibrahim[1] build a reference ontology and then profile scholars' background knowledge based on this ontology. Segura[15] uses

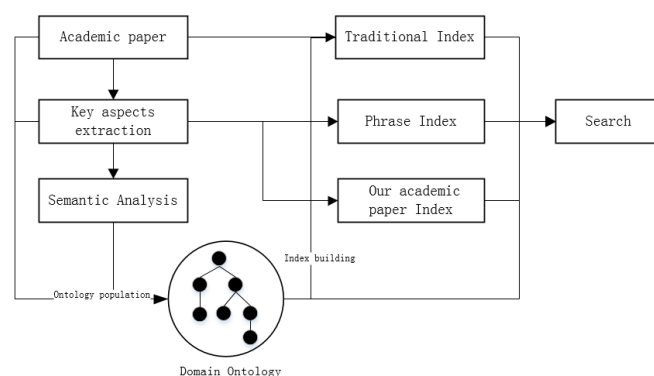


Fig. 1. The overall architecture of literature search framework

terminology relationships in the ontology for query construction. However, directly run over the ontology individuals will cause heavy read operations[9][13] which would not be scaled to large knowledge bases. A scalable alternative approach is to transform the data in ontology repository into a special inverted index structure[11]. As we also add articles and their extracted and analyzed contents as ontology instances. To avoid this scalability problem, what we do is to transform the data in ontology repository into inverted index.

The contribution of this paper is to propose a literature search framework which supports a novel perspective of retrieving research point and implementation method of articles. Meanwhile, this framework offers the facilities to researchers to recognize the key aspects of research contents in articles easily without reading them.

The rest of this paper is organized as follows: Section II introduces the details of components of our literature search framework. Section III presents the evaluation of our framework and the application prototype. Finally, conclusions of this paper is given in Section IV.

II. ARCHITECTURE OF LITERATURE SEARCH FRAMEWORK

The overall architecture of our literature search framework is shown in Fig. 1 which contains four components: key aspect extraction, semantic analysis, ontology population, and index building. Firstly, academic papers are collected and analyzed, key aspects are identified and extracted. Secondly, semantic

TABLE I. SOME RULES FOR EXTRACTION PHRASES

research point	implementation method
prep_on(focus,NP VP)	dobj(using,NP)
dobj(solve,NP VP)	prep_on(based,NP)
dobj(improve,NP VP)	dobj(apply,NP)
prep_on(work,NP VP)	nn(model,NP)
dobj(present,NP VP)	dobj(extend,NP)

analysis is done on the phrases to calculate the semantic similarity of phrase pair, and cluster the phrases that have the similar meaning. Thirdly, the result obtained in the previous two components are mapped into ontology instances. Finally, the index is built by transforming ontology instances into an inverted index structure.

A. Key aspects extraction

Key aspect extraction is used to extract research point and implementation method from articles. Among them, research point is described by noun phrases or verb-object phrases, implementation method is described by noun phrases. The phrases we extract are the input of semantic similarity component.

As the most appropriate description of research point and implementation method often appear in the abstract and introduction of articles, and often have the fixed expressions, like the sentence started with “we study”, “we focus on”, etc. We use the method proposed by Gupta[8] who extract phrases from these kind of sentences by dependency tree[4]. Considering the phrases they extract contain much noise, and lack of semantic analysis, phrase-structure tree is used to restrict the extracted phrases to noun phrases and verb-object phrases, and semantic analysis is done on these phrases.

TABLE I shows a part of manual rules we use to extract phrases according to dependency tree and phrase-structure tree. These rules are all have the form $d(T, POST)$, where T represent the trigger word (like “using”, “focus”) we set, d is the relation name (like “dobj” and “prep_on”) according to dependency tree, and $POST$ is the pos tagging of phrases to be extracted according to phrase-structure tree. If the trigger word has a successor whose dependency is d , we extract the corresponding noun phrases or verb-object phrases from the parser-tree, and label them with one of the aspects(category).

However, using these seed rules to extract phrases is far from enough, we use bootstrapping method to form new $d(T, POST)$. If the extracted phrase is a sub-tree of a sentence dependency tree, the ancestor (parent or grandparent) could be T , the dependency between T and the head of sub-tree could be d , the pos tagging of the sub-tree could be $POST$. For a set of candidate rules generated by phrases, each of them has a score:

$$\omega(c_m) = \frac{|p, \forall p \in P_m|}{\sum_{i=1}^n |p, \forall p \in P_i|} \quad (1)$$

where c_m represent the candidate rules for category m , p represent the phrase to form c_m , denominator represent the total numbers of phrase p occurs in the corpus, and numerator represent the number of phrase p occurs in the corresponding category. While all the candidate rules which do not exist in the previous iterations are assigned scores, we select the top 2 rules in each category respectively.

B. Semantic analysis

This step is to analyze the phrases obtained in key aspect components by semantic similarity, and cluster the phrases which have the similar meaning. In fact, merging the phrases that have the similar meaning to a canonical form is difficult, because different authors have different ways of expression, and there could exist ambiguity problems. We use semantic similarity to analyze the phrases. Firstly, C-value method[5] is used as a filter to extract the key terms from phrases which are obtained from key aspect extraction. Secondly, we analyze the phrases by calculating the semantic similarity of key terms in them. That is, using semantic-vector[12] approach to compute phrase similarity. We simplify this measure by only using term similarity scores as weights.

$$Sem_{sim}(T_1, T_2) = \tanh\left(\frac{Overlapping(T_1, T_2)}{Length(T_1) + Length(T_2)}\right) \quad (2)$$

where $Sem_{sim}(T_1, T_2)$ represents the semantic similarity between two terms T_1 and T_2 , which is based on Lesk algorithm. $Overlapping(T_1, T_2)$ means the number of shared terms of Wiki T_1 and T_2 pages. $Length(T_1)$ means the total terms appears in Wiki T_1 pages. This equation means if one term frequently appears in the Wiki body page of the other term, or both always appear in the same Wiki page, they could be similar.

Based on the phrase similarity, phrases that have the similar meaning could be clustered by Chameleon algorithm. Here we use the first phrase to represent phrases that are clustered together.

C. Ontology population

This step is to enrich the ontology with the obtained semantic data which are used for semantic reasoning in the next step. We have designed an domain ontology which is related to information retrieval in advance, and we only focus on the sub domain, semantic retrieval and ontology of this ontology. Therefore, our ontology population is a process to map the semantic data obtained in the above components into ontology instances. For example, from a narration of research point “acquire user-profiles through the construction of user ontologies”, we obtain a semantic canonical form “ontology-based user profiles acquisition”, and “manual user ontology” is used as the implementation for this research point. Using these data and relations, we create OWL individuals for each article. In order to not affect the recall performance, we still create empty properties “Unknown” if no information about key aspects is extracted from the article in terms of the key aspects. Certainly, ontology population is not restricted with these kinds of data, the basic information of articles, like authors, publisher are also included. These information are also added as OWL individuals to the ontology for each of the articles if they do not exist in the knowledge base.

D. Index building

The academic paper index in the literature search framework is built for facilitating the paper representation and retrieval which could avoid the scalability problem of querying the ontology directly. So we build the indexing mechanism with extracted and inferred information using domain ontology.

TABLE II. INDEX STRUCTURE

Field	Description
Title	The title of an article
Author	Authors of an article
Keywords	Keywords in an article which are given by authors themselves
Focus	Research point of the article
FocusCenter	The represented phrases of research points
Method	Implementation method of the article
MethodCenter	The represented phrases of implementation methods
Narration	Original full-text description

```
(?article pre:hasFocus ?focus)
(?focus pre:canonicalForm ?focusCenter)
->(?article pre:hasFocusCenter ?focusCenter)
```

Fig. 2. Inference rules

1) *Index structure*: Index structure has utmost importance in meeting users needs. The index structure not only contains traditional metadata such as title and keywords, but also some semantic fields. We also add original full-text description associated with each semantic field. This is important if the content in research point or implementation fields is unknown, since the full-text description can ensures the recall value of traditional full-text search at least. The index structure and the description of each field are shown in TABLE II.

2) *Inference and rules*: The content of semantic fields in index structure is not filled simply by some basic property. They keep the inferred information according to the inference rules. We give an example of inferring the FOCUS field, which is shown in Fig.2. If an article A *hasFocus* B, and B's *canonicalForm* is C, we can infer A *hasFocusCenter* C. Using these rules, we are able to refine the research point of articles. Note that Method field is also filled using semantic rules like this.

III. FRAMEWORK EVALUATION AND DEMONSTRATION

A. Framework evaluation

We apply this literature search framework to computer science domain, and considering that we are familiar with semantic retrieval sub domain, we crawled 303 articles using keywords *semantic* and *ontology* from ScienceDirect whose generic metadata are stored in ontology and manually label them for testing. Besides, we set up a series of search queries, which cover the search for research point and implementation method(shown in Table III). Q1,Q2,Q3 and Q4 are used to verify the effectiveness of semantic searching for research point, Q5 is used to verify the effectiveness of searching for research point and implementation method. Q6 is used to verify the effectiveness of searching for implementation method. For comparison, we build the following three kind of Lucene indices for detailed evaluation.

- 1) **Traditional index** contains basic information available in ontology of articles, such as title, author, keywords and abstract[10]. It's used to illustrate the importance of extraction.

TABLE III. EVALUATION QUERIES

Q1	Find the articles who is focus on knowledge representation (query:"knowledge representation")
Q2	Find the articles whose research point is to build an ontology (query:"build ontology")
Q3	Find the articles whose focus is on semantic similarity (query:"semantic similarity")
Q4	Find the articles who solve semantic ambiguity or heterogeneous problem (query:"semantic ambiguity or heterogeneous")
Q5	Find the article who use dynamic ranking algorithm to improve the retrieval performance (query:"retrieval performance, dynamic ranking")
Q6	Find the article who use WordNet to study research points. (query:"WordNet")

- 2) **Phrase index** contains the phrases of research point and implementation method we extract besides the basic information. It's used to illustrate the importance of semantic analysis.
- 3) **Our academic paper index** (shown in Table II) contains semantic information in addition to the phrase index.

we observe the top-10 results for each query, which are generally most concerned by the search users. We use well-known evaluation metrics Average Precision (AP) and Precision (Eq.3) at 10 (P@10) to measure the performance for each query. Average Precision (Eq.4) is a metric to measure the overall effect in terms of precision, recall and ranking (we use the default ranking mechanism in Lucene).

$$precision = \frac{\text{number of retrieved relevant documents}}{\text{total number of retrieved documents}} \quad (3)$$

$$AP = \frac{\sum_{k=1}^{|Q|} P(k) \times rel(k)}{|RA|} \quad (4)$$

where $|Q|$ is the number of queries, $|RA|$ is number of relevant documents, $P(k)$ is the precision at cut-off k , and $rel(k)$ is 1 if the k th result is relevant, zero otherwise.

Table IV and Table V show that, compared with "Traditional index" and "Phrase index", the academic paper index of our framework has some improvement on AP and Precision respectively for six representative queries. The reason is that authors could use a variety of expressions to describe their research point or implementation method. "Traditional index" and "Phrase index" are not able to recognize the narration as they do not contain the key words. However, as our framework concerns the real meaning of a phrase, it can do an expansion of the query, and returns the relevant articles.

Although our index performs better on most of the queries, Q6 is an exception. We hypothesize that the semantic similarity component used in our framework is not suitable for all the terms. If "WordNet" appears in a phrase, the other terms beside "WordNet" will be also involved in similarity process, and thus leads to an inaccurate result.

B. Demonstration

We apply this literature search framework to support the retrieval from the perspective of research point. Fig.3 shows the result when user's query is "ontology". This framework

TABLE IV. QUALITY OF RESULTS BY AP.

Query	Traditional index	Phrase index	Our academic paper index
Q1	0.14	0.17	0.22
Q2	0.07	0.08	0.25
Q3	0.23	0.29	0.29
Q4	0.19	0.35	0.41
Q5	0.25	0.32	0.32
Q6	0.25	0.55	0.44
Mean	0.19	0.29	0.32

TABLE V. QUALITY OF RESULTS BY P@10.

Query	Traditional index	Phrase index	Our academic paper index
Q1	0.5	0.7	0.8
Q2	0.5	0.5	0.9
Q3	0.6	0.7	0.7
Q4	0.4	0.6	0.7
Q5	0.8	0.8	0.8
Q6	0.5	0.9	0.9
Mean	0.55	0.7	0.8

firstly return the research point, this could make users know all research points related to the query. When users choose a research point, our framework would show the articles related to the research point, which is shown in Fig.4. Users could also directly view the research point and implementation method of articles without reading them carefully, which offers the facilities to users.

IV. CONCLUSION

The main goal of our work is to help researchers quickly move to unfamiliar fields. In order to achieve this goal, we propose a literature search framework by analyzing research point and implementation method of articles. To realize this framework, we use dependency tree and phrase-structure tree by matching extraction rules in sentences, and semantic analysis is done on the phrases, then build the index after ontology population. Its application to academic retrieval shows that our approach can outperform the traditional search easily. Moreover, we demonstrate the prototype of this framework, it offers a novel perspective retrieval of research point, and facilities users to recognize key aspects of research contents in articles easily.

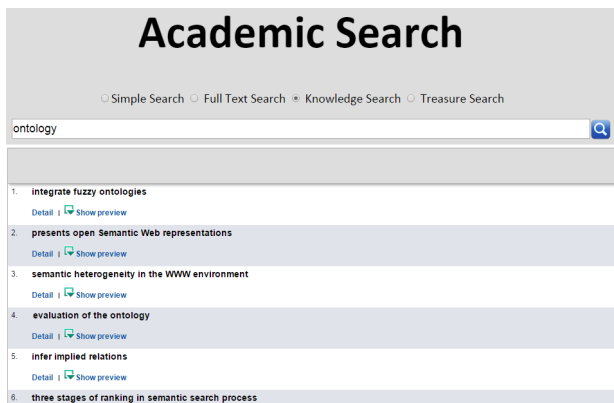


Fig. 3. Research point related to the query “ontology”



Fig. 4. Articles related to the research point “ontology-based similarity measures”

V. ACKNOWLEDGE

This work is supported by the National Natural Science Foundation of China (No. 61370137) and the 111 Project of Beijing Institute of Technology.

REFERENCES

- [1] Amini, B., Ibrahim, R., Othman, M.S., Nematbakhsh, M.A.: A reference ontology for profiling scholars background knowledge in recommender systems. *Expert Systems with Applications* 42(2), 913–928 (2015)
- [2] David Hall, Daniel Jurafsky, C.D.M.: Studying the history of ideas using topic models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2008)
- [3] David M. Blei, Andrew Y. Ng, M.I.J.J.L.: Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 2003 (2003)
- [4] De Marneffe, M.C., MacCartney, B., Manning, C.D., et al.: Generating typed dependency parses from phrase structure parses. In: *Proceedings of LREC*. vol. 6, pp. 449–454 (2006)
- [5] Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries* 3(2), 115–130 (2000)
- [6] Fukuda, S., Nanba, H., Takezawa, T.: Extraction and visualization of technical trend information from research papers and patents. *D-Lib Magazine* 18(7), 6 (2012)
- [7] Gerrish, S., Blei, D.M.: A language-based approach to measuring scholarly impact. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. pp. 375–382 (2010)
- [8] Gupta, S., Manning, C.D.: Analyzing the dynamics of research by extracting key aspects of scientific papers. In: *IJCNLP*. pp. 1–9 (2011)
- [9] Haofen Wang, Kang Zhang, Q.L.T.T.Y.Y.: Q2Semantic: A Lightweight Keyword Interface to Semantic Search. *The Semantic Web: Research and Applications* (2008)
- [10] Jianhua, S.: strategy analysis of library selection of resource discovery system. *Information Science*. 6 (2015)
- [11] Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N.K., Alpaslan, F.N.: An ontology-based retrieval system using semantic indexing. *Information Systems* 37(4), 294–305 (2012)
- [12] Li, Y., McLean, D., Bandar, Z.A., O’shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on* 18(8), 1138–1150 (2006)
- [13] Ma, S., Tian, L.: Ontology-based semantic retrieval for mechanical design knowledge. *International Journal of Computer Integrated Manufacturing* 28(2), 226–238 (2015)
- [14] Peter Dolog, W.N.: Semantic web technologies for the adaptive web. *The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science* (2007)
- [15] Segura, N.A., García-Barriocanal, E., Prieto, M., et al.: An empirical analysis of ontology-based query expansion for learning resource searches using merlot and the gene ontology. *Knowledge-Based Systems* 24(1), 119–133 (2011)

- [16] Tateisi, Y., Shidahara, Y., Miyao, Y., Aizawa, A.: Annotation of computer science papers for semantic relation extraction. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (2014)



Zhuyue Si was born in Shandong province, China, in 1991. In 2009, she received the Bachelor's Degree in College of telecommunication, Tongji University, Shanghai, China. And now, she is working toward the Master Degree of School of Computer Science, Beijing Institute of Technology, Beijing, China, specializing in data mining and information retrieval.



Gaokun Dai was born in Sichuan province, China, in 1992. In 2010, he received the Bachelor's Degree in Beijing Institute of Technology Beijing, China. And now, he is pursuing Master Degree in School of Computer Science, Beijing Institute of Technology, Beijing, China, specializing in information retrieval and text automatic summarization.



Zhendong Niu received the Ph.D. degree in computer science from the Beijing Institute of Technology, Beijing China, in 1995. He was a postdoctoral researcher at the University of Pittsburgh, Pittsburgh, PA, USA, from 1996 to 1998, a research/adjunct faculty member at Carnegie Mellon University, Pittsburgh, PA, USA, from 1999 to 2004, and a joint research professor in the Information School at the University of Pittsburgh from 2006. He is a Professor and the Deputy Dean of the School of Computer Science and Technology at the Beijing Institute of

Technology. His research areas include informational retrieval, software architecture, digital libraries, Web-based learning techniques, etc. Dr. Niu was the recipient of the IBM Faculty Innovation Award in 2005 and the New Century Excellent Talents in University of Ministry of Education of China in 2006.