

## Literature retrieval based on citation context

Shengbo Liu · Chaomei Chen · Kun Ding · Bo Wang · Kan Xu · Yuan Lin

Received: 17 November 2013 / Published online: 25 January 2014  
© Akadémiai Kiadó, Budapest, Hungary 2014

**Abstract** While the citation context of a reference may provide detailed and direct information about the nature of a citation, few studies have specifically addressed the role of this information in retrieving relevant documents from the literature primarily due to the lack of full text databases. In this paper, we design a retrieval system based on full texts in the PubMed Central database. We constructed two modules in the retrieval system. One is a reference retrieval module based on citation contexts. Another is a citation context retrieval module for searching the citation contexts of a specific paper. The results of comparisons show that the reference retrieval module performed better than Google Scholar and PubMed database in terms of finding proper references based on topic words extracted from citation context. It also performed very well on searching highly cited papers and classic papers. The citation context retrieval module visualizes the topics of citation contexts as tag clouds and classifies citation contexts based on cue words in citation contexts.

---

S. Liu · K. Ding · B. Wang (✉)  
WISElab, Dalian University of Technology, No. 2, Linggong Road, Ganjingzi District,  
Dalian 116024, China  
e-mail: bowang1121@gmail.com

S. Liu  
e-mail: liushengbo1121@gmail.com

K. Ding  
e-mail: dingk@dlut.edu.cn

C. Chen  
College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia,  
PA 19104-2875, USA  
e-mail: chaomei.chen@drexel.edu

K. Xu · Y. Lin  
Information Retrieval Laboratory, Dalian University of Technology, Dalian, China  
e-mail: xukan@dlut.edu.cn

Y. Lin  
e-mail: zhlin@dlut.edu.cn

**Keywords** Literature retrieval · Citation context · Tag cloud · Citation context classification

## Introduction

Literature retrieval is concerned with searching the most relevant bibliographic information. When writing a paper, researchers have to find some papers as the intellectual base of their work. These papers should be the most relevant papers not only to the subject of the paper in discussion but also to the sub-topics of the paper. Normally, researchers will search the relevant papers on the web. But the great amount of scientific information being published makes it difficult for users to identify the most relevant information. For example, in the biomedical domain alone, around 1,800 new papers are published each day (Hunter and Cohen 2006).

With the development of the field of scientometrics, citations are often used in literature retrieval to improve the retrieval efficiency. Four types of citations can be applied to enhance the performance of literature retrieval. The first type is citation count which is treated as an indicator for ranking the retrieval results, and finding the most cited papers. Bibliographic coupling and co-citation measures are another two types based on citation linkages to find the most relevant papers. Bibliographic coupling refers to a linkage between two documents which have one or more identical references (Kessler 1963), whereas co-citation is defined as a linkage between two documents concurrently cited by another document (Small 1973). These two types of citations can be used to reveal the relationships between documents. Some examples have showed that they could improve the performance of information retrieval (Eto 2012; Nanba et al. 2000; Pao 1993; Small 1973). Many popular literature search engines, such as CiteSeer<sup>1</sup> and Google Scholar<sup>2</sup> also use the links between articles and documents provided by citations to enhance their ranked retrieval results. The fourth type of citations is the citation context. The citation context of a given reference can be defined as the sentences that contain a citation of the reference. For instance, the sentence “This comparison is made using BLASTX (Nanba and Okumura 1999)” is a citation context of the reference (Nanba and Okumura 1999). One may also define a citation context based on more sentences before and/or after the citation sentence. Many researchers have tried to enhance search performance by incorporating citation context into information retrieval systems (Bradshaw 2003; Mercer and Marco 2004; Nakov and Hearst 2004; O’ Connor 1982).

Actually, citation context can provide direct information about an instance of citation. Researchers did not use these citation contexts directly to retrieve literature, but use these citation contexts to improve the traditional retrieval systems. One of the most important reasons is that it is very hard to collect all the citation contexts of the retrieved literatures. In the past, information regarding citation context was not readily accessible due to the lack of full text of citing papers. Researchers often had to extract the necessary information through a manual process. For example, O’Connor (O’ Connor 1982, 1983) extracted single words from citation contexts. Small (1986) extracted concepts from citation contexts to name a cluster of a co-citation network. In recent years, full text literatures are more

<sup>1</sup> Scientific Literature Digital Library, <http://citeseer.ist.psu.edu>.

<sup>2</sup> Google search engine, for peer-reviewed scholarly literature, <http://scholar.google.com>.

accessible. PubMed Central provides full text documents in XML format. In this paper we will introduce the design of a literature retrieval system based on all full text documents in PubMed Central.

We design two modules for the retrieval system. One is the reference retrieval module based on citation context. In this module, the system can be used to recommend relevant literatures to users. Another is the citation context retrieval module for searching the citation contexts of a specific paper. This module facilitates users to analyze citation contexts to further understanding the retrieved literatures. We expect that this system could help researchers find the needed documents more quickly and accurately.

## Related work

### Citation context analysis

The citation context analysis includes the application of the citation position and citation content.

Citation positions are considered in co-citation analysis. Elkiss et al. (2008) and Liu and Chen (2012) studied co-citations in an article at four levels: the sentence level, the paragraph level, the section level, and the paper level. Elkiss found that papers co-cited at a finer granularity are more similar to each other than papers co-cited at a coarser granularity. For example, papers co-cited at the sentence level have a stronger relationship than papers co-cited at the section level. Liu found that sentence-level co-citations are potentially more efficient candidates for co-citation analysis. Gipp and Beel (2009) classified the co-citation into five categories based on occurrence positions: within the same sentence, the same paragraph, the same chapter, the same journal and the same journal but different editions. In each category, a co-citation is given a different value of 1, 1/2, 1/4, 1/8 or 1/16. The result shows that the weighted co-citation analysis yields much more similar documents than traditional co-citation analysis. Callahan et al. (2010) used a similar method to calculate the co-citation strength. Recently, Boyack et al. (2012) used the co-citation proximity to improve the co-citation clustering performance. He found that taking into account reference proximity from full text can increase the textual coherence of a co-citation cluster solution by up to 30 % over the traditional approach based on bibliographic information.

Citation content can be used to identify the nature of a citation. The attributions and functions of a cited paper can be identified from the semantics of the contextual sentences (Siddharthan and Teufel 2007). Nanba and Okumura (1999, 2005) collected citation context information from multiple papers cited by the same paper and generated a summary of the paper based on this citation context information. They also extracted citing sentences from citation contexts and generated a review. Mei and Zhai (2008) and Mohammad et al. (2009) found that the summarization of citation contexts is very different from the abstract of the cited reference. Nakov et al. (2004) referred to citation contexts as citances—a set of sentences that surround a particular citation. Citances can be used in abstract summarization and other natural language processing (NLP) tasks such as corpora comparison, entity recognition, and relation extraction. Small (1979) studied the context of co-citation and analyzed the context in which the co-citation paper was mentioned. In addition, he analyzed the sentiment of the co-citation context (Small 2011). Mei (2008) defined the length of citing sentences as 5, 2 sentences before the citation and three sentences after. In this study, we use the sentence with the citation tag as the citation context.

Anderson and Sun (2010) analyzed the citation contexts of a classic paper in organizational learning which was published by Walsh and Ungson in the *Academy of Management Review*. The results provided a richer understanding of which knowledge claims made by Walsh and Ungson have been retrieved and have had the greatest impact on later work in the area of organizational memory, and also what criticisms have been leveled against their claims. Our research also designed a module for searching citation contexts of any specific paper. It is very helpful for researchers to understand the citation value of a reference.

#### Citation context used in citation retrieval

O'Connor (1982, 1983) assumed that citing statements give some information about the cited document. Cue words were extracted from the citation context and applied as index terms for the cited document. Then these index terms were used to improve the search effectiveness. Bradshaw (2003) proposed a reference directed indexing (RDI) method to improve information retrieval of scientific literature. RDI also used similar method to O'Connor's to create index terms from citation contexts. RDI considered both the relevance of a document to the query terms and the number of papers citing it.

Mercer and Di Marco also described their work on using citances to improve indexing tools for biomedical literature (Mercer and Marco 2004). The first step of their work was using cue phrases in citances to predefine the citation classification. Then they applied these classifications to improve existing citation indexes. Ritchie (2008) take the explicit, content words from citation contexts and index them as part of the cited document. And the results showed that the citation-enhanced document representation increases retrieval effectiveness across a range of standard retrieval models and evaluation measures.

Our reference retrieval module is similar to RDI, but we directly use the citation context as the retrieval field and rank result according to frequency of references which are corresponding to the citation contexts. The advantage of this approach is that the citation context could reveal the citation value of a reference.

## Data and method

Our procedure consists of four major components: (1) data collection, (2) citation context extraction, (3) index creation, and (4) retrieval system design (See Fig. 1).

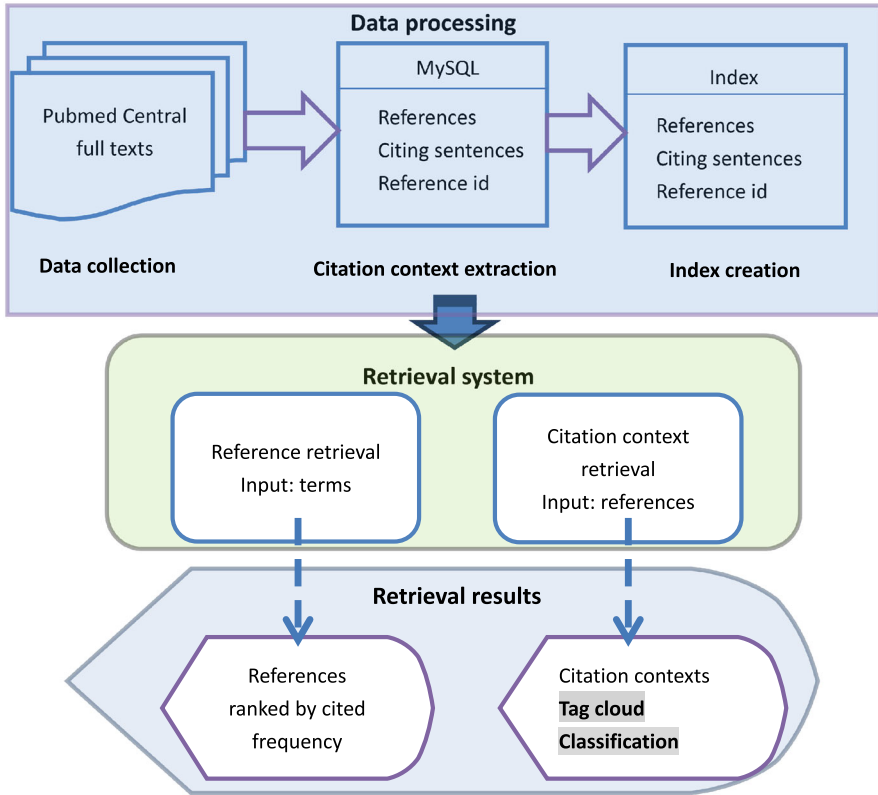
### Data collection

All full text papers in PubMed Central were selected in this research. The data was downloaded on July 23 2012. There are 3,431 journals with 622,801 papers. All of these papers and their references were used to build the database for citation retrieval.

Papers published on December 2012 in *BMC Bioinformatics* were chosen as the test dataset. There are 26 papers and 751 citation contexts.

### Citation context extraction

The full text literatures in PubMed Central are XML files. Figure 2 shows an example of a XML file with reference information. The citation context and its corresponding reference information are extracted and saved in MySQL database. In this paper, citation context is



**Fig. 1** The citation retrieval system design

defined as one citing sentence with the reference tag. 17,551,920 citing sentences were extracted from 622,801 papers.

**Index creation**

The aim of creating an index is to speed up the retrieval. Although citing sentences are stored in MySQL, the retrieval speed is very slow due to the large size of the citation context dataset. Therefore, indexing is necessary in this research. Lucene v3.5 is employed to create indexes for the retrieval field of citation context and cited reference. Not all the words in citation context are indexed, the stop words will be filtered out automatically during indexing.

**Retrieval system design**

The system includes two modules. One is the reference retrieval module; the other is the citation context retrieval module.

Content:

... ..

The method is based on extending a Bayesian unsupervised classification method for MLST data [[4](#)] to a semi-supervised setting, where the existing curated MLST database plays the role of training data.

... ..

References:

... ..

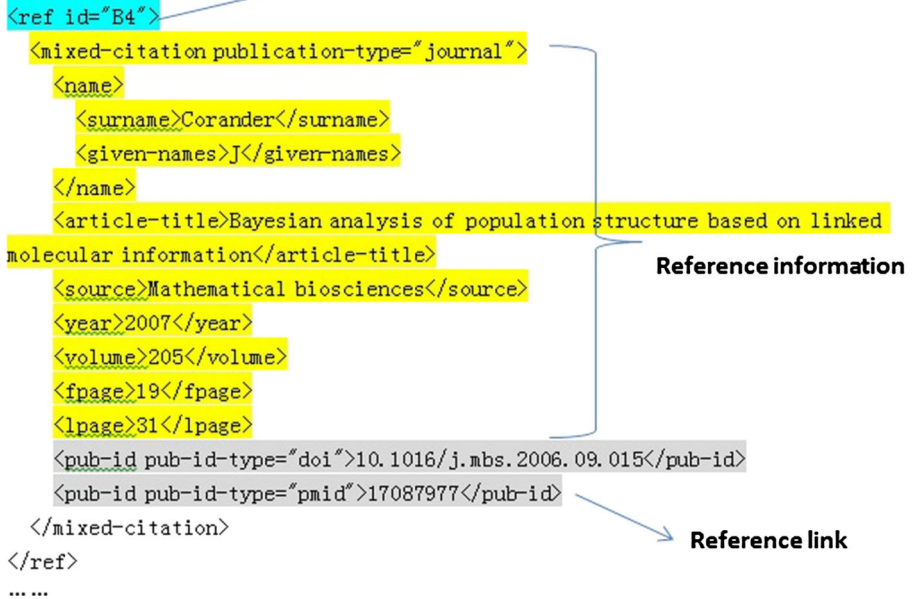


Fig. 2 Extracting citation context from XML files

Reference retrieval module

In this module, the retrieval field is the citation context. The indexes of 17,551,920 citation contexts have been created. Researchers use topic terms to search the relevant citation contexts. But the citation contexts are not the final results. The references corresponding to these citation contexts are the results that researchers want to get. Each citation context corresponds to one or more references. The results will be ranked by corresponding counts of the citation context. Each retrieved reference has a unique reference link to the title and abstract of the reference. Figure 3 shows an example of retrieval references related to “lung cancer”. “Parkin DM, 2005, CA Cancer J Clin, V55, P74” ranked first in the results. It was cited by 55 sentences, which means that “Parkin DM, 2005, CA Cancer J Clin, V55, P74” was cited 55 times on the topic of “lung cancer”. The general information about this paper can be found through the linkage. “Parkin DM, 2005, CA Cancer J Clin, V55, P74” might also have been cited numerous of times on other topics. The citation context retrieval module which we discussed later provides the total cited times and topics of a chosen reference.

**Fig. 3** An example of reference retrieval

### *Citation context retrieval module*

The retrieval field of this module is the reference field. Researchers could use author, year, and/or journal information to find target references. The results show the citation frequency and citation contexts of the references. One reference could have one hundred citation contexts or even more. It is time consuming to read all these citation contexts. So we further analyzed citation contexts from two aspects. One is a topic analysis; the other is a classification of citation context. Tag cloud is employed to represent the citation contexts with topic terms in the topic analysis. Tag cloud (word cloud or text cloud) is a visual representation for text data, typically used to depict keyword metadata (tags) on websites, or to visualize free-form text. Tags are usually single words, and the importance of each tag is shown with font size or color (Halvey and Keane 2007). An example is showed in Fig. 4. The reference “Parkin DM, 2005, CA Cancer J Clin, V55, P74” is used in this example which is the one we used in the reference retrieval module. 554 citation contexts have been retrieved. The reference retrieval module has retrieved 55 of 554 citation contexts related to “lung cancer”. The other citation topics of this reference were represented in tag cloud. Figure 5 shows the tags cloud of the citation contexts with single words. The main citation topic of this reference is the common causes of cancer death. The citation subtopics involve different kinds of cancer, different countries and genders that cancer occurs. Lung cancer is just one aspect of the citation topics.

A tag cloud could give us a more intuitive summary of which part the content of a paper has been cited. But we do not know the motivation of the citer. When the citer cited a paper, did he attempt to praise the work or criticize some drawbacks? Such motivation information will be very helpful for the user to comprehend the impact of a cited paper. We design a classification function to classify citation contexts based on the motivation of citers. Normally, semantic analysis is used to analyze the sentiment of a sentence in Natural Language Processing. But there are few sentiment words in scientific papers. It is hard to appraise the citation context with semantic analysis method (Verlic et al. 2008). Therefore, we choose to use the cue words method in a similar way as it has been used in recent work, notably by Small (2011) and Teufel et al. (2006).

**PMC Reference Search Engine**

Parkin DM, 2005, CA Cancer J Clin, V55, P74

**1. Parkin DM, 2005, CA Cancer J Clin, V55, P74**  
 Cited by 554 Sentences, and cited by 461 papers. Q=1.2017354

1. Breast cancer is the most common malignancy diagnosed in women across the world today.1 Treatment options for breast cancer patients include surgery, radiotherapy, chemotherapy, and hormone replacement therapy.
2. Hepatocellular carcinoma (HCC) comprises the third highest cause of cancer death worldwide.1 Surgical resection offers a cure, however most patients are ineligible for resection at the time of diagnosis because of tumor extent and underlying liver disease.
3. Approximately half of new gastric cancer cases occur in East Asia, including 41% in China and 11% in Japan.1,2
4. There are an estimated 493,000 new cases of cervical cancer that result in 233,000 deaths per year worldwide [15].
5. HCC is the third most lethal neoplasm causing an estimated 600,000 deaths annually due to late diagnosis underlying liver disease and lack of effective treatment options (Parkin et al., 2005; Llovet and Bruix, 2008; Bruix and Sherman, 2005; Llovet et al., 2008).
6. Head and neck squamous cell carcinoma (HNSCC) is a malignancy that involves cancers of the lip, oral cavity, pharynx, hypopharynx, larynx, nose, nasal sinuses, neck, ears and salivary glands (Parkin, et al., 2005).
7. Nearly 40 to 60% of HNSCC patients subsequently develop chemoresistance, locoregional recurrences or distant metastases (Jemal, et al. 2008, Parkin, et al. 2005).
8. These tumor cells display persistent cell growth leading to chemoresistance and recurrence (Parkin, et al., 2005).
9. Although the incidence of gastric cancer has been declining in the past several decades, it remains the second most common cause of cancer death (Parkin et al. 2005).
10. Colorectal cancer (CRC) is a major cause of cancer mortality worldwide, with approximately 500,000 recorded deaths from the disease in 2002 [1].
11. GLOBOCAN data from the International Agency for Research on Cancer showed that, in 2002, cancer of the kidney accounted for 1.9% of all new cases of cancer diagnosed and was responsible for 102,000 deaths worldwide [16].
12. Globally, lung cancer remains the most common cancer and the leading cause of cancer-related mortality [1].
13. Nevertheless 411 000 annual deaths were reported, being the leading cause of cancer mortality in women [1].
14. Gastric and esophageal cancers are the second and sixth leading causes of cancer related death worldwide, respectively [15].
15. The estimated worldwide incidence is about 500,000 new cases per year [1], with over 63,000 events/year in Europe [2].
16. The majority (over 90%) of these new cases are solid tumors, and each year over 12 million people die from solid tumors (1,2).

**Fig. 4** An example of citation context retrieval



**Fig. 5** Tag cloud of citation contexts

Following the work of Spiegel-Rösing (1977) and Teufel et al. (2006), citation contexts will be classified into three categories as positive, negative and neutral. Table 1 shows the description of each category. The positive category has three sub categories and the negative category has two sub categories. Table 2 lists some cue word instances corresponding to



**Table 1** The description of each category

Category	Description
Positive	
1	Affirm or praise the cited work
2	Apply the methods, tools or databases of the cited paper
3	Comparison of methods and results
Negative	
1	Point out the weakness of the citation
2	Contain negative cue words
Neutral	
1	Contain no cue words

**Table 2** The cue words of each category

Category	Subject	Cue words
Positive		
1	Citation	Best, significant, crucial, fundamental.....
2	Citer	Use, utilize, employ, apply, be based on.....
3	Citer	Superior to, better, similar to, consistent with.....
Negative		
1	Citation	Failed, limited, lack.....
2	Citation	Not, although, but.....
Neutral		
1	Citation	Describe, discuss, publish, use, apply.....

each category. The subject of each sentence is also needed during the classification. The sentences “We use this tool...” and “They use this tool...” correspond to different categories. The passive voice will be converted to the active voice before classifying.

The classification function is next to the “cloud” button (see Fig. 4). When clicking on the “classify” button, the classification results will be shown to the user. For the reference “Parkin DM, 2005, CA Cancer J Clin, V55, P74”, there are 25 positive citation contexts, 529 neutral citation contexts, and no negative citation context. This reference is about the study of global cancer statistics, so most of the citation contexts are neutral.

**Result testing**

*Reference retrieval testing*

In order to check the performance of the retrieval system, 26 new papers with 751 citation contexts from *BMC Bioinformatics* were collected. The topics of each citation context were identified with 1–4 topic words manually. For example, the sentence “As a feature of reaction rules, some techniques focus on physicochemical properties and structures (Small 1973)” will be tagged with “physicochemical”, “properties”, and “structures”. These topic words are used as retrieval terms to search references. Not all the citation contexts have topic words: for example, “It evolves the two different populations within the context of each other (Kessler 1963; Mei and Zhai 2008)”. The citation topic of this reference

might have been expressed in the sentences before or after this citation context. The dataset was divided into four groups by time period, in order to check the influence of time. We chose 50 citation contexts which have explicit topic words for each period. The papers published earlier tend to receive more citations. So we expect that the retrieval system will perform better on the early time period. If the corresponding reference of a citation context appears among the top ten retrieval results, we mark this retrieval as a successful retrieval. Otherwise, we mark it as an unsuccessful retrieval.

Two comparative studies are designed using Google Scholar and PubMed. Google Scholar is one of the most popular search engines for researchers. It could retrieve literature from full text and rank the results not only by relevance but also related to citation frequency. PubMed database is a specialized database of Biomedical. The data source of this paper is PubMed Central which is a subset of PubMed, PubMed database has been chosen as one of the testing experiments as well.

We adopt the same retrieval strategy with the retrieval system designed in this paper. For Google Scholar, the retrieval results are sorted by relevance. If the corresponding reference of a citation context appears among the top ten retrieval results, this retrieval will be marked as a successful retrieval. Otherwise, it is an unsuccessful retrieval. For PubMed database, we retrieve the topic words in the field of title and abstract. The database just provide one result ranking method which is ranking by publish year. So if the corresponding reference of a citation context appears among the retrieval results, we mark this retrieval as a successful retrieval.

### *Citation context classification testing*

Although the cue words are selected based on large amounts of statistical data, the rules in Table 2 are basically artificially defined, the accuracy needs to be verified. This experiment will compare the differences between cue word method and the manual judgment method. Firstly, 1,000 citation contexts are randomly selected from MySQL database and divide these citation contexts into ten groups. Each group has 100 citation contexts. Secondly, these citation contexts are classified by domain experts, both citation contexts and the texts surrounding them are provided, they can get as much information as they need from the paper. This classification result will be treated as a standard classification. Then we adopt the cue word method and the manual judgment method to classify these data sets. In the manual judgment method, a judge, who did not participate in the standard classification procedure, classifies the citation contexts based only on cited sentences. Finally, *t*-test is employed to test the significance of the difference between the two methods. Ideally the classification results of these two methods shall have no significant differences.

## **Results**

### Reference retrieval testing results

The testing result of the reference retrieval module is shown in Table 3. The testing data was separated into four time periods based on the number of references in each year. The four periods are 1973–2000, 2001–2005, 2006–2008, and 2009–2011. The results show that the retrieval system performs very well for the early time period with the accuracy rate of 68 % which is higher than the CRM-crosscontext method performs (He et al. 2010). The CRM-crosscontext is a citation recommendation method with the precision 42 %. For the

**Table 3** Retrieval performance of the retrieval system

	1973–2000	2001–2005	2006–2008	2009–2011	Total
Successful	34	30	30	19	113
Unsuccessful	16	20	20	31	87
Accuracy rate	68 %	60 %	60 %	38 %	56.5 %

period 2001–2005 and 2006–2008, the accuracy rates are the same. They both have reached 60 % which is a little lower than 1973–2000. For the most recent time period, the system did not perform very well. The accuracy rate of this period is only 38 % which is the lowest in the four time periods.

Table 4 shows ten instances of the successfully retrieved topics and references. The topics are extracted from citation contexts and the original references that the citation contexts used are ranked the first in all retrieval results respectively. Most of these successfully retrieved topics are about tools and methods. The highly cited conclusions could also be retrieved successfully. For example, “Han JD, 2004, Nature, V430, P88” is retrieved on topic “data party hubs”. This paper was cited 100 times on this topic.

Although some of the citation contexts with explicit topics were not retrieved successfully, it did not mean that the retrieval system does not fit for these topics. Table 5 shows three examples of comparisons of the original references with the recommended references retrieved from our system on the same topics. The testing dataset used “Chang CC, 2011, ACM Trans. Intell. Syst. Technol, V2” as the reference of topic “LIBSVM”. But our system recommended another paper of Chang’s which was published in 2001 and received 34 citations on topic “LIBSVM”. For topic “BLAST e-value”, the original reference was Karlin’s paper which had just one citation on this topic. The recommended reference had been cited 66 times on this topic. It is hard to judge which reference is better. It is impossible to read all the related articles while we are conducting our research. Our recommended references are retrieved based on the behavior of all other authors. Our system definitely has some value which cannot be ignored.

Tables 6 and 7 are the testing results in Google Scholar and PubMed. The average successful rates are 44 and 13 % respectively, which are lower than the retrieval system designed in this paper. We find that there are two reasons for the low successful rate of PubMed. One is that the numerous of conference papers in the references are not indexed by PubMed. Another reason is that the retrieval fields are title and abstract only, which could not provide enough information for searching topic words.

For the test of Google Scholar, the accuracy rates are lower than our retrieval system in the early three periods. But in the time period of 2009–2011, the performance is obviously better than our retrieval system. Our retrieval system has the lowest accuracy rate in the time period of 2009–2011 because of the lower citation frequency of the references in this time period. In Google Scholar, the search is not only related to the citation frequency, but also related to the topic words and full texts. So new theories and methods could be easily retrieved in Google Scholar.

According to Tables 3 and 6, the successful instances in our system and Google Scholar are 113 and 88. But there are just 63 instances can be successful retrieved in both our system and Google Scholar. 50 of 113 successful retrieved instances in our system could not be retrieved in Google Scholar. And 25 of 88 successful retrieved instances in Google Scholar could not be retrieved in our system.

**Table 4** 10 instances of successful retrieved topics

Topics	References	Freq
Weblogo	Crooks GE, 2004, Genome research, V14, P1188	376
Date party hubs	Han JD, 2004, Nature, V430, P88	100
BiMax	Prelic A, 2006, Bioinformatics, V22, P1122	40
PredictNLS	Cokol M, 2000, EMBO Rep, V1, P411	20
SVMLight	Joachims T, 1999, Making large-scale SVM learning practical	11
Bron-Kerbosch algorithm	Bron C, 1973, Commun ACM, V16, P575	10
Amino acid compositions	Hua S, 2001, Bioinformatics, V17, P721	7
PMSprune	Davila J, 2007, TCBB, V4, P544	6
APBioNet	Tan TW, 2010, BMC Genomics, V11, PS27	5
Chemical tagger	Hawizy L, 2011, J Cheminf, V3, P17	4

**Table 5** Comparison of original references and retrieved references

Topics	Sources	References	Freq
LIBSVM	Original	Chang CC, 2011, ACM Trans. Intell. Syst. Technol, V2	3
	Retrieved	Chang CC, 2001, LIBSVM: a library for support vector machines	34
Graphviz	Original	Ellson J, 2001, Lecture Notes in Computer Science Springer-Verlag, P483	0
	Retrieved	Ellson J, 2003, Graph Drawing Software, P127	5
BLAST e-value	Original	Karlin S, 1990, Proceedings of the National Academy of Sciences of the United States of America, V87, P2264	1
	Retrieved	Altschul SF, 1990, J Mol Biol, V215, P403	66

**Table 6** Testing results in Google Scholar

	1973–2000	2001–2005	2006–2008	2009–2011	Total
Successful	21	13	24	30	88
Unsuccessful	29	37	26	20	112
Accuracy	42 %	26 %	48 %	60 %	44 %

**Table 7** Testing results in PubMed database

	1973–2000	2001–2005	2006–2008	2009–2011	Total
Successful	8	4	5	9	26
Unsuccessful	42	46	45	41	174
Accuracy	16 %	8 %	10 %	18 %	13 %

## Classification results

Table 8 shows the citation context classification testing results with the cue word method and the manual judgment method. Each number in this table represents the degree of

**Table 8** Testing results in PubMed database

Group	1	2	3	4	5	6	7	8	9	10
Cue word method	96	98	96	95	99	99	96	96	98	96
Manual judgment method	98	99	100	98	99	98	99	99	100	100

consistency with standard classification, for example, 96 citations contexts were classified into the same category with standard classification according to the cue word method, while 98 according to the manual judgment method in this same group 1. As shown in the table, there is no significant difference from standard classification. Cue-word-based classification has an average of 96.9 % consistency with the standard classification, while the percentage for the manual judgment is 99 %.

To further illustrate the significance of difference between cue word method and the manual judgment method, a *t* test is used. *t* test method is used here for verification since the small size of the sample, result shows that the cue word method and the manual judgment method only have a difference value of 0.001 under the 95 % confidence interval. Thus we consider the cue word method is reliable for citation context nature evaluation.

### Discussion

The retrieval system designed in this paper is based on the large amount of full text papers in PubMed Central. Most of the databases do not provide the full texts. Therefore, the retrieval system in this paper is particularly suitable for the field of biomedicine. With the development of information science, the citation retrieval system will extend to other fields where full text databases are available.

The reference retrieval module shows its effectiveness on searching papers published early and papers with high citation frequencies which is what we expected. It is also very effective in retrieving papers that regarding introduce methods or tools. The reference retrieval module will perform better on retrieving basic or classic papers in a specified field. But papers with lower citation frequencies will be hard to find in this system, since the retrieval field of this module is citation context. In comparing with Google Scholar, some of the references not found in Google Scholar can be retrieved in our system, and vice versa. We expect a combination of the two searching methods would increase the overall performance.

The citation context retrieval module provides all the citation contexts of a specific reference. These citation contexts may contain many topics. Tag cloud is employed to represent these topics. Classification is introduced to characterize the nature of citation contexts and citers' motivations. The topics of the citation contexts greatly enhance the meaning of a reference. The retrieval results enrich our understanding of which knowledge claims by the references have been used and have had the greatest impact on subsequent work, and also what criticisms have been leveled against their claims. They also can be used to evaluate the impact of a reference together with the citation frequency.

There are some limitations in our research. The reference retrieval module was designed based on the citation context. If one paper was not cited, it could not be found in this system. The retrieval field of the reference retrieval module is citation context. If the citation contexts of a reference do not contain topic words, they will not be retrieved. Although tag cloud method could identify the main topic words of the retrieved citation contexts, these topic words still need to be clustering.

A test version of the literature retrieval system is available on the World Wide Web at: <http://ir.dlut.edu.cn:8090/PMCSEARCH/>.

## Conclusion

We designed a literature retrieval system based on citation contexts extracted from full text publications in biomedicine. The reference retrieval module is for searching publications which have been cited on topics related to the querying terms. The citation context retrieval module is for searching the citation contexts of a specific paper and for visualizing the contributions of the specific paper in a tag cloud. The results showed that this retrieval system was particularly accurate in retrieving highly cited papers and classic papers, whereas the accuracy was reduced when searching less cited papers and newly published papers. The performance of our retrieval system is better than Google Scholar and PubMed database in our testing experiment. The citation context retrieval module could identify different citation topics of a reference and classify the citation contexts. In summary, our work demonstrates the potential of using citation contexts in enhancing the retrieval of scientific publications and improving our understanding of the impact of a specific publication on subsequent work.

**Acknowledgments** This research is supported by National Natural Science Foundation of China (grant number 61272370), the specialized research fund for doctoral tutor (20110041110034), and the Fundamental Research Funds for the Central Universities. Part of the research was conducted during Shengbo Liu's visiting doctoral studentship at the iSchool at Drexel University.

## References

- Anderson, M. H., & Sun, P. Y. T. (2010). What have scholars retrieved from Walsh and Ungson (1991)? A citation context study. *Management Learning*, *41*(2), 131–145.
- Boyack, K. W., Small, H., & Klavans, R. (2012). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, *64*, 1759–1767.
- Bradshaw, S. (2003). *Reference directed indexing: Redeeming relevance for subject search in citation indexes*. Paper presented at the Proceedings of the 7th European conference on digital libraries, Trondheim.
- Callahan, A., Hockema, S., & Eysenbach, G. (2010). Contextual cocitation: Augmenting cocitation analysis and its applications. *Journal of the American Society for Information Science and Technology*, *61*(6), 1130–1143.
- Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, *59*(1), 51–62.
- Eto, M. (2012). Evaluations of context-based co-citation searching. *Scientometrics*, *94*(2), 651–673.
- Gipp, B., & Beel, J. (2009). *Identifying related documents for research paper recommender by CPA and COA*. Paper presented at the Proceedings of International Conference on Education and Information Technology, Berkeley.
- Halvey, M., & Keane, K. (2007). *An Assessment of Tag Presentation Techniques*. Paper presented at the 16th International World Wide Web Conference, Banff.
- He, Q., Pei, J., & Kifer, D. (2010). *Context-aware Citation Recommendation*. Paper presented at the 19th International World Wide Web Conference, Raleigh.
- Hunter, L., & Cohen, K. (2006). Biomedical language processing: What's beyond pubmed? *Molecular Cell*, *21*(5), 589–594.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25.
- Liu, S., & Chen, C. (2012). The proximity of co-citation. *Scientometrics*, *91*(2), 495–511.

- Mei, Q., & Zhai, C. (2008). *Generating impact-based summaries for scientific literature*. Paper presented at the Proceedings of ACL '08, Columbus.
- Mercer, R. E., & Marco, CD. (2004). *A design methodology for a biomedical literature indexing tool using the rhetoric of science*. Paper presented at the BioLink workshop in conjunction with NAACL/HLT, Boston.
- Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishnan, P., Qazvinian, V., Radev, D., & Zajic, D. (2009). *Using citations to generate surveys of scientific paradigms*. Paper presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder.
- Nakov, P. I., Schwartz, A.S., & Hearst, M.A. (2004). *Citances: Citation sentences for semantic analysis of bioscience text*. Paper presented at the SIGIR 2004 Workshop on Search and Discovery in Bioinformatics, Sheffield.
- Nanba, H., Kando, N., & Okumura, M. (2000). *Classification of research papers using citation links and citation types: Towards automatic review article generation*. Paper presented at the Proceedings of the American society for information science, Chicago.
- Nanba, H., & Okumura, M. (1999). *Towards multi-paper summarization using reference information*. Paper presented at the The 16th International Joint Conference on Artificial Intelligence, Stockholm.
- Nanba, H., & Okumura, M. (2005). *Automatic detection of survey articles*. Paper presented at the The Research and Advanced Technology for Digital Libraries, Berlin.
- O' Connor, J. (1982). Citing statements: Computer recognition and use to improve retrieval. *Information Processing and Management*, 18(3), 125–131.
- O' Connor, J. (1983). Biomedical citing statements: Computer recognition and use to aid full-text retrieval. *Information Processing and Management*, 19(6), 361–368.
- Pao, M. L. (1993). Term and citation retrieval: A field study. *Information Processing and Management*, 29(1), 95–112.
- Ritchie, A. (2008). *Citation context analysis for information retrieval*. New Hall: University of Cambridge.
- Siddharthan, A., Teufel, S. (2007). *Whose idea was this, and why does it matter? Attributing scientific work to citations*. Paper presented at the Proceedings of NAACL/HLT-07, Rochester.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science and Technology*, 24(4), 265–269.
- Small, H. (1979). *Co-citation context analysis: The relationship between bibliometric structure and knowledge*. Paper presented at the Proceedings of the ASIS Annual Meeting, Medford.
- Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. *Journal of the American Society for Information Science*, 37(3), 97–110.
- Small, H. (2011a). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, 87(2), 373–388.
- Small, H. (2011b). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, 87(2), 373–388.
- Spiegel-Rösing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science*, 7, 97–113.
- Teufel, S., Siddharthan, A., & Tidhar, D. (2006). *Automatic classification of citation function*. Paper presented at the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.
- Verlic, M., Stiglic, G., Kocbek, S., & Kokol, P. (2008). *Sentiment in Science-A Case Study of CBMS Contributions in Years 2003 to 2007*. Paper presented at the Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on Parallel Processing.