# Submission recommendation system based on Internet of things engine

Peng Wang and Chuansheng Wu

University of Science and Technology Liaoning, China

gykwcs@163.com

## Abstract

**In this work, the proposed orchestrating and sharing system for online paper aims at managing papers from information collecting, paper editing, paper typesetting, paper submitting to paper sharing. In the five aspects above, there are many available tools which help science researchers write papers, but these tools work separately not cooperatively. Orchestrating and sharing system for online paper integrates functions of these tools, which offers one-stop service. As an important part of this system, the recommendation for paper submission is to provide valuable information about the latest international conferences and journal for paper publication. When papers are written, our system, a context-aware solution for paper, automatically obtains the keywords from context. Given that the recommendation for paper submission is subject-oriented search, we design a recommendation system for paper submission based on vertical search engine, which enhances the search accuracy by the improved URL-based filtering algorithm and the improved content-based filtering algorithm.**

## Keywords

**Search engine; vertical search; paper submission; PageRank algorithm.**

## 1. Introduction

PapersCloud[1] is an online paper orchestrating and sharing system that supports the whole life cycle of science papers (briefly called paper in this paper below). The basic meaning of the life cycle can be popularly understood as "the whole process from the

cradle to the grave" (Cradle-to-Grave). In accordance with the definition of the life cycle, we propose a paper life-cycle approach. Paper life cycle is the process that from collection information, paper editing, paper type-setting, paper submission to paper sharing. In our system, paper life cycle includes references recommendation, paper editing, paper type-setting, paper submission recommendation and paper sharing management. References recommendation is to offer some related papers for the interesting topics. Paper editing and type-setting get the input of the paper and typesetting according to a certain format. Paper submission recommendation gives a list of institutes for submitting the paper for publication. The last is to manage the available papers remotely. This paper mainly aims at papers submission recommendation.

Paper submission recommendation gives information about paper publication for the authors after their accomplishment of their papers. How to always get better result is the main part of this paper. We can get the research field and the keywords from the paper, and we search the most related information for the author from our recommendation library. According to the preference of every author, we give out the result in different ways: for new authors we give out the most related and for old users we just give out the recently information such as the main dates of their preferences.

## 2. System Overview

This system is mainly designed to provide information of conferences and periodicals when users want to submit their papers. And the system continues to use the classical framework, it is consists of web spider module, dumper module, index module and query module. The main process of the system is shown as following figure 1. For instance, when user finishes a paper about cloud computing, our system will automatically obtains the keywords from context. When system obtains the keywords,

web spiders crawl the web pages form Internet, after that we filter some useless pages and becoming the web page database. Then web page go through information extraction, calculation of the PageRank value ect. and after these steps, data can be indexing. The last step is similarity calculation, and the query result will show to the user.

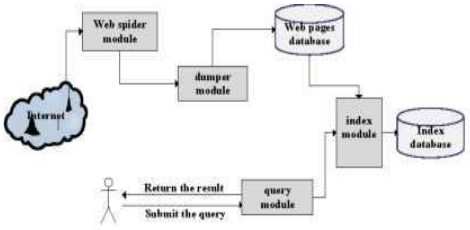


**Fig. 1** the main process of the system

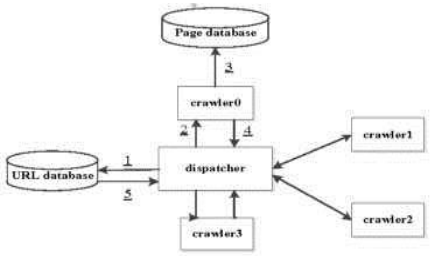


**Fig.2** the crawlers working process

The first module of search engine is web spider module. It is the foundation of the search engine. It is in charge of downloading useful web pages from the World Wide Web. All search data are derived from the work of web spider module.

This is a huge project because there are tens of thousands of web pages [2] on the Internet. World Wide Web has bow tie structure. Because of this, we select directory type website as the crawler start page. Figure 2 shows how crawlers work.

In figure 3, we can see crawler get URL from the URL database, then download the web page and after the filter useless pages form the page database.

In this module, our main consideration is the selection of the URL library. The choice of text search engine is portals and directory-type site. As for this, it get much more information about sub-sites or related sites of the portal. Thus, there is much more unconcerned information it get. Because the selection of URL library influences the search result to a large extent and we pay more attention on the search for international conferences and journals, we use both method of URL setting and portals for the crawler. To set the certain sites, we can get more accurate information than we get from the portals, on the contrary, we can get entire information from the method of portals.

The second module is dumper module. For dumper module, it basic and primary work is categorized extract[3] valuable information from the semi-structured page. And this information can represent the attributes of the page, such as anchor text, title, and content.

Because the information we get through the crawler are complicated and uncertain, we need to eliminate the unconcerned information and put the rest to our recommendation library. As to dumper module, it dose the work to search results according to the conditions of the authors such as deadline of paper submission, journals or conferences, SCI or EI and the rest from our recommendation library.

The third one is index module. The index module is the search engine's data warehouse, it storages and indexing millions of thousands of web pages. The purpose of indexing the web page is convenient for the query in the next stage. We are need to

indexing the page crawl from the web that can accelerate the speed of query.

In this module, we will first be in progress the Chinese word segmentation, it is mainly to segment the sentence into a collection with suitable word. Then, we will calculate the PageRank value. The

result of offline calculation will be returned a list of PageRank, including a PageRank value of every page. And will be easily retrieved in the query module. At last, we will index the pages.

The last module is query module. Query module directly faces the users. It receives the query request by online users, and gives the user result in accordance with the calculation by retrieving, sorting and abstract extract and so on.

Figure 3 shows the process of query module. Firstly, the system receives the query request from the user, and then compares the request with the cache hit. If the request in it, we directly show the result to the user, and if not we query the word from index module. When the index module returns the result, rearrangement the file and extract the abstract, form the result page to user. The whole query requirements not only faster, but also are able to provide users with available results.

## 3. Detailed Design

Our system provide the function that can offer some information about recommended conferences or periodicals for users, which requires the system to consider how to control the search results that will not be offset and filter the useless query information. These issues will be exhaustively described in next context.

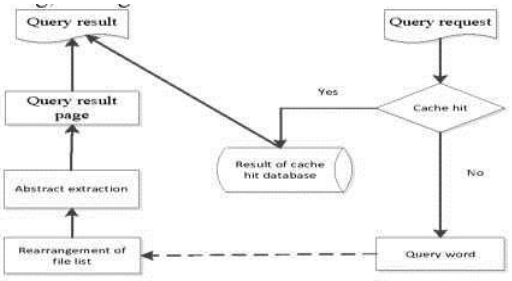


**Fig.3** the process of query module

### 3.1 URL Filter Analysis and Implementation

Search engine crawlers work mechanism is priority to grab the web pages which has the high relevant to the subject. Web pages were sorted according to the page ranking, and only keep themes that above the URL threshold.

Currently widely used URL filtering algorithm is consists of two classes, PageRank algorithm and HITS algorithm [4]. The basic idea of PageRank algorithm [5-6] is that, web pages from a number of high quality web links, must be have the high quality web pages. HITS algorithm bases on the idea that the really value of the page is highly relevant to the theme of the user's search content. HITS algorithm easily occurs that pages deviate from the core theme and irrelevant results returned.

Compared with the two algorithms, PageRank algorithm is in dominant position. So, we selected the PageRank algorithm and improved it. PageRank algorithm is simple, but it ignores the user's understanding of web page. It gives different web pages the same weight, this will search high value but have little relationship about the subject.

In order to improve the results, we improved PageRank algorithm by using the user's visiting navigation path diagram to modify the traditional PageRank value.

Different web page has different probability to be visited by the users, so you can through the web page's visiting probability to express the initial PageRank value. The simple way is (1). Ap is the number that page p was visited; ^Ap is the number that all pages were visited.

We should consider the actual situation that different users access to different pages, and unbalanced to set the current page's ability to recommend the outlink web page. And combined with actual visiting condition, the PageRank value expresses as below.

In formula (2), Wp(Ti) is the weight that from page Ti visiting page A , and this weight is proportional to the number page Ti outlink page A, inverse proportion to the page Ti outlink all the pages. So Wp(Ti) is:

In formula (3), AT is the number that user through page Ti visited page P; A^^
is the number that user through page Ti all outlinks, 5(T) is the page TJs all outlinks.
According to the formula (1), the PageRank value expresses is:

$$PR_1(P) = \frac{A_P}{\sum A_P} \quad (1)$$

$$PR_{n-1}(P) = (1-d)*1+d*\sum \frac{PR_{n-1}(T_i)*W_p(T_i)}{C(T_i)} \quad (2)$$

$$W_p(T_i) - \frac{A_P^T}{A_{B(T_i)}} \quad (3)$$

$$PR_{n-1}(P) = (1-d)*W_p+d*\sum \frac{PR_{n-1}(T_i)*W_p(T_i)}{C(T_i)} \quad (4)$$

$$W_P = \frac{A'_P}{A_P} \quad (5)$$

$$PR_{n-1}(P) = (1-d)*\frac{A'_P}{A_P}+d*\sum \frac{PR_{n-1}(T_i)*W_p(T_i)}{C(T_i)} \quad (6)$$

In formula (4), Wp is the weight that user random visit some page, the weight is proportional to the number that user not through other link visit page A, inverse proportion to the number that user visit page A, the expresses is

In formula (5), A'p is the number that user not through other link visit page A, Ap is

the number that user visit page A.

At last, we got the modified PageRank value:

By improving the PageRank algorithm, the URL's filtering accuracy is improved. And through that we can get more useful page that related to the user's requirements.

### 3.2 Content Filter Analysis and Implementation

Content filtering need to make sure that the content filtering algorithm with contextueel and real-time result, therefore, filtration precision and filtration velocity becomes a key content filtering criterion. Current algorithms for matching model include Boolean model, vector space model and analysis semantic model. Boolean model is a strict matching model, its fast speed, very convenient to realize and suitable for structured information. Vector space model [7] is expressed as a vector to page document, and it will be submitted to customers with the search content. Semantic analysis model based on keyword matching will search for the link between the search item and the actual content to build for a semantic model.

In content-based filtering algorithm, we need to evaluate the similar degree between page and the search subject, that is to say the keywords[8] and the web page will be sort by correlation. We would improve the algorithm to increase the search relevance of content. Specific ideas of the improve algorithm:

Combined the keywords in the text with the frequency and location to determine the relative weights

Web crawler is to analyze and collect the network data that have higher relative weights, and filter irrelevant information

Using vector space model to collected text for the N dimensional vector, and calculate the similarity

Therefore, we simplified the user's query keyword and network resources to a vector which means weight. The method of vector space model works as figure 4.

Vector similarity algorithm is as follows: Suppose we have two text data which have relation to the author's paper, and expressed as D1, D2. wik, w2k indicate text weight. Similarity sim(D 1, D2) can be expressed with the distance between the vector:

Through the improved algorithm can further filter out content which are not related to web pages or the related degree is not high, increase the precision of the system.

## 4. Experiment Studying

In this section, we describe an experiment to test our vertical search engine. We crawl about 26,756 web pages on the Internet by open source search engine Nutch. After that we transported the crawling web pages to database. The keywords, "cloud computing", "data base", "mobile computing" respectively are chosen for the experiments. And we compute the accuracy. The comparison of the two results as follow figure 5. Obviously the accuracy of the improved one is higher to the original one.

$$sim(D_1, D_2) = \frac{\sum_{k=1}^{n} w_{1k} * w_{2k}}{\sqrt{\sum_{k=1}^{n} w_{1k}^2 \sum_{k=1}^{n} w_{2k}^2}} \tag{7}$$

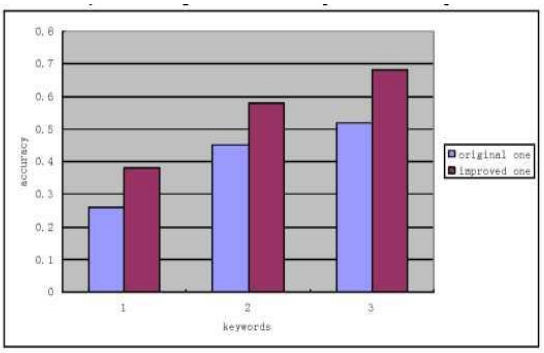


Fig. 5 the contrast result

## 5. Conclusion

In this paper, one vertical search engine for paper submitted was designed, which can give users a help to search available conference or periodicals. We improved the accuracy through the selection of URL library, data filtering, better PageRank and better similarity algorithm. The key problem, in detailed design, filtering algorithm was discussed in detail. With the improved algorithm, the search engine's query result is comprehensive and its precision is high and it gives users more available results.

There are also some further problems to solve, such as better user interface to display the result, more reasonable Chinese words segmentation and site revisit time and so on. All above mentioned issues deserve further research.

## References

[1] L. Zhang, L. Yang, Z. Xu. papers cloud: a composing-free, collaborative editing platform for scientific papers [C]. "The Proceeding of the 2012 IET International Conference on Frontier Computing-Theory, Technologies and Applications (IETFC 2012)," ISBN 978-1-849-19604-8, pp, 286-291(2016).

[2] J.C. Yang, P.L. Ling."Improvement of PageRank Algorithm for Search Engine." Computer Engineering. 35, 35-37(2015)

[3] M. Chau, H.C. Chen."A Machine learning approach to web page filtering using content and structure analysis," Elsevier Science Direct. 44(2), 482-494(2016)

[4] J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment, "Journal of the ACM (JACM). 46(5), 604-632 (1999)

[5] T. Hjghjf, Topic-sensitive PageRank. "Proceedings of the 11th International Conference on World Wide Web (WWW02), Honolulu, "Hawaii. pp, 517-526(2016)

[6] L. Page, S. Brin, R. Motwani. “The Page Rank Citation Rnking: Bringing Order to the Web.” Standford Digital Library Technologies Project 1998 [Page, et al.1998] (2013)
[7] Y.M. Zhang, J.F. Zhou. “A train able method for extracting chinese entity names and their relations. ”Proceedings of the Second Chinese Language Processing Workshop. 12, 66-72 (2016)
[8] C. Amit, K. Jaewoo. “Selective approach to handing topic oriented tasks on the World Wide Web. ”Proceeding of the 2017 IEEE Symposium on Computational intellingence and Data Ming (CIDM2007), ISBN 1-4244-0705-2/07. pp, 343-348 (2016)