http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (1): 164-167, 2011
ISSN 1812-5638

Research on Information Fusion Model for Patent Retrieval

B. Jin, L. Feng, H.J. Piao and Z.Q. Diao
School of Innovation Experiment, Dalian University of Technology,
Dalian, Liaoning 116024, People's Republic of China

Abstract: Most of the present patent retrieval systems only use a single retrieval engine and cannot meet the growing demand of patent uses. Fusion of the results set of various retrieval engines is one of the best solutions for patent retrieval. In this study, we propose a Result Set Relevance Inference (RSRI) model for patent retrieval fusion. It makes use of two result sets returned from keyword retrieval and semantic retrieval engine and then fuses them to a new result set after calculation as per their weights. Using an experimental data set, results showed that the RSRI model can improve the recall and mean average precision in most cases.

Key words: Patent retrieval, information fusion, results merging

INTRODUCTION

Patent retrieval is widely used for scientific research and intellectual property protection. But current patent retrieval technology lacks the accuracy and has difficulty in dealing with the multi-sources information. Information fusion is one of the key issues for patent retrieval. In this study, we introduced a novel patent retrieval fusion approach.

The World Intellectual Property Organization (WIPO) has predicted that judicious use of patent information could, to a reasonable extent, lead to the prevention of duplication of research, which could save as much as 60% in time and 40% in funding (Jiang *et al*., 1999). On the basis of an estimate by WIPO, patent publications cover approximately 90-95% of the results of scientific research worldwide, probably greater than the percentage covered by all scientific journals. Because patent authorities periodically publish patent data for public inquiry, enormous data have been accumulated for many years. However, the tools and concepts for retrieval used in patent research are not on par with those in scientific research.

In this study, we propose a Result Set Relevance Inference (RSRI) model to improve the performance of patent retrieval. Firstly, we apply diverse retrieval techniques in patent retrieval system, such as keyword retrieval and semantic retrieval. Then, we fuse the results from those retrieval methods. Finally, we give out more accurate results to meet the requirements of users.

How to make good use of patent is very important for small- and medium-sized manufacturing enterprises (Jin *et al*., 2008) and case-based projects (Jin and Teng, 2007). Xu *et al*. (2010) has developed a web service based system with patent information for manufacturing enterprises. These works need more effective patent retrieval technologies. Jin *et al*. (2007) studied patent mining method basing on sememe statistics and key-phrase extraction. With this method, the sememe and key-phrase are extracted from patent for semantic retrieval and key-phrase retrieval. Jin *et al*. (2010) proposed a novel textual chunk retrieval technology for patent.

Currently, various solutions have been suggested for merging separate result lists obtained from distributed retrieval engines. As a first approach and taking only the rank of the retrieved items into account, the results might be interleaved in a round-robin fashion. According to previous studies, such interleaving schemes have a retrieval effectiveness of around 20-40% below that achieved from single retrieval models, working with a single huge collection representing an entire set of documents.

In order to account for document scores computed for each retrieved item, Fox and Shaw (1999) have proposed a series of fusion strategies, such as CombMIN, CombMAX, CombSUM and CombMNZ. Belkin *et al*. (1995) has proved through experiment that the effect of multiple retrieval engines is better than a single retrieval engine. Hu (2006) has proved that the result, fused from keywords and semantic retrieval engines, is better than either one.

Xu and Croft (1999) suggested that documents could be gathered according to their topics and a language model associated with each topic. Callan *et al*. (1995) have presented a Collection Retrieval Inference network (CORI) which considers each collection as a single gigantic

Corresponding Author: L. Feng, School of Innovation Experiment, Dalian University of Technology, Dalian, Liaoning 116024, People's Republic of China Tel: +86-411-84707445 Fax: +86-411-84708599

document. Ranking the collections is similar to document ranking methods used in conventional information retrieval system.

Rasolofo *et al.* (2001) have proposed using result length to calculate merging score (LMS) which is to CORI. It uses a weight to adjust the combination score of documents. Regarding the score of information set corresponding to the document, if it is higher than the average, the document's combination score is increased. On the contrary, the combination score of the document is reduced.

RESULT SET RELEVANCE INFERENCE MODEL

In this study, RSRI restricts that all retrieval engines apply the same relevance calculation method. It hypothesizes that documents, which repeatedly appear in each retrieval engine's result set, are related. Thus, fusing the result sets of different retrieval engines is feasible. One of the core problems of patent information retrieval fusion is weight estimation. According to our problem, we determine the weight of the retrieval engine as:

$$w_i = w_{il} \times w_{is} \quad (1)$$

where, w_{il} is the weight according to the length of the result set and w_{is} is the weight according to the relevance of the result set.

Callan *et al.* (1995) have presented a Collection Retrieval Inference network (CORI) which considers each collection as a single gigantic document. Inference network is a probabilistic approach to information retrieval. In CORI, the weight of how one might weigh results from different collections is defined as (Callan *et al.*, 1995):

$$w = 1 + |C| \cdot \frac{s - \bar{s}}{\bar{s}} \quad (2)$$

where, |C| is the number of collections searched; s is the collection's score and $\bar{s}$ is the mean of the collection scores.

Rasolofo *et al.* (2001) improved the CORI and determined the collection score as:

$$s_i = \log\left[1 + \frac{l_i \cdot K}{\sum_{j=1}^{|C|} I_j}\right] \quad (3)$$

where, K is a constant (set to 600) and l_i is the number of documents retrieved by the i-th collection.

The above weight can be used in our model as w_{il} in Eq. 1. The other weight w_{is} is defined as:

$$w_{is} = 1 + \left[(s_{is} - \bar{s}_s)/\bar{s}_s\right] \quad (4)$$

where, $\bar{s}_s$ is the average value of all the scores of the retrieval engines and s_{is} is the score of retrieval engine c_i, which is calculated according to the relevance of the repeated documents.

$$s_{is} = \log\left[1 + \frac{\bar{s}_{is}}{\sum_{j=1}^{|C|} \bar{s}_{js}}\right] \quad (5)$$

where, $\bar{s}_{is}$ is the average value of all the relevance scores of the repeated documents in the result set of retrieval engine c_i.

Finally, we provide the document relevance score as:

$$s_i = \sum_{j=1}^{|C|} w_j s_{ij} \quad (6)$$

where, s_{ij} is the score of document s_i graded by retrieval engine c_j. If document s_i does not exist in the result set of retrieval engine c_j, it takes 0 for s_{ij}.

RESULTS MERGING STRATEGIES

After defining the RSRI model for patent information retrieval fusion, lists of results returned from collections can be combined into a final single ranked list. To resolve this problem of combination, a merging strategy has been suggested, shown as Fig.1.

Step 1: Determine the retrieval elements from patent claims

Step 2: Retrieve patent using various retrieval engines, such as keyword retrieval and semantic retrieval

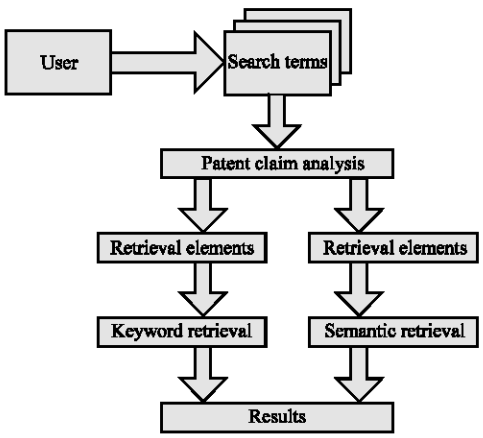

Fig. 1: Merging strategies

Step 3: Fuse the results with our RSRI model and then return the final result to the user.

RESULTS

We randomly selected 12000 patent documents from the State Intellectual Office (SIPO) of People's Republic of China as our test data set. The patent documents included 7 classes, such as mechanics, electronics, aerostatics, chemistry, computer, physics and agriculture.

Evaluation: We carried out an empirical evaluation of our fusion model using patent documents from the State Intellectual Property Office (SIPO) of People's Republic of China. Our goals were to assess the overall effectiveness of our fusion model.

We measured fusion quality by quality metrics such as recall, precision and mean average precision. These metrics have long been used to assess the quality of literature searches (Liu *et al*., 2008). In this study, recall is the proportion of correct answers that are returned. Precision is the proportion of returned patents that are correct answers. In this study, we use the precision of the first 30 patents (p@30) to describe the proportion of the related patents in the first 30 patents of the result set. It is helpful to examine the distribution of the related patents. Mean Average Precision (MAP) is the distribution of the related patents in the result set:

$$M = \left(\sum_{j=1}^{k} \frac{j}{P_j}\right) \Big/ k \quad (7)$$

where, k is the quantity of related patents retrieved by some retrieval engine and P_j is the position of the related patents in the result set.

Experimental results: In this study, we compared several models with our RSRI model for patent retrieval, such as keyword retrieval, semantic retrieval, CombMNZ model and LMS model, the experimental results are shown in Table 1.

As shown in Table 1, the recall and precision of the RSRI model are higher than those of single retrieval models, such as keyword retrieval and semantic retrieval. The reason is that keyword retrieval is based on the part of speech of the retrieval key elements, while semantic retrieval is based on the semantic of the retrieval key elements. In Chinese, there are many phenomena of polysemy. Thus, the two single retrieval models are mutually complementary and independent, not the relation of inclusion. According to the results, we can prove that the fusion model has been markedly heightened. It is feasible to apply the retrieval fusion model to Chinese patent retrieval.

We choose the CombMNZ and LMS model for comparison with our RSRI model. The CombMNZ is a kind of classic fusion method which has a very simple algorithm and good fusion effect. LMS gives each retrieval engine a weight and makes the algorithm easily realized in the Web. As shown in Table 1, our RSRI model gains improvement. The advantage of retrieval fusion is discussed as follows:

- When the precisions of keyword retrieval and semantic retrieval are about 0.9, as shown in type 7 in Table 1, the difference among the three fusion models is not obvious. The reason is that there are lots of related patent documents in the result sets. The fusion process is just to resort the result sets
- When one precision of the two single retrieval models is about 0.9, as shown in type 1 and 5 in Table 1, the RSRI model is better than the LMS method. In the LMS model, as more documents are retrieved, more related documents are contained in the result set. But when the semantic retrieval engine returns more documents, the keyword retrieval engine cannot return more related documents. Our RSRI model improved the shortcoming of the LMS
- When one precision of the two single retrieval models is between 0.5 and 0.8, while the other is no larger than 0.9, as shown in type 2, 3 and 6 in Table 1, the RSRI model gains the best. This is most normal in actual retrieval. Our RSRI model has taken all the factors into account, such as the length of result set and related documents

Table 1: Single retrieval engine's precision and five kinds of result sets' P@30

	Recall			P@30					MAP				
Type	Key.	Sem.	RSRI	Key.	Sem.	Comb	LMS	RSRI	Key.	Sem.	Comb	LMS	RSRI
1	0.320	0.340	0.340	0.533	0.000	0.533	0.400	0.400	14.394	2.424	14.270	8.199	8.399
2	0.180	0.180	0.180	0.300	0.300	0.300	0.300	0.300	9.000	4.517	8.900	8.582	9.000
3	0.120	0.240	0.240	0.200	0.400	0.400	0.400	0.400	5.857	7.810	9.691	10.733	11.843
4	0.320	0.240	0.360	0.533	0.400	0.500	0.467	0.533	6.251	7.810	14.008	13.542	13.502
5	0.120	0.420	0.480	0.200	0.633	0.767	0.700	0.733	6.000	17.561	21.471	20.549	21.070
6	0.420	0.460	0.460	0.700	0.733	0.767	0.767	0.767	21.000	19.878	22.533	24.653	24.653
7	0.200	0.580	0.580	0.333	0.967	0.967	0.967	0.967	10.000	29.000	29.000	29.000	29.000

- When both precisions of the two single retrieval models are less than 0.5, as shown in type 4 in Table 1, the RSRI model is no better than the other fusion models. The reason is that there are lots of unrelated patent documents in the result sets. The simplest model might get the best results

ACKNOWLEDGMENTS

This study was supported by National Natural Science Foundation of People's Republic of China (Grant No. 60773213), Program for New Century Excellent Talents in University (Grant No. NCET-09-0251), Natural Science Foundation of Liaoning Province (Grant No. 20071092) and Scientific Research Fund of Liaoning Provincial Education Department (Grant No. 2010035).

CONCLUSIONS

In this study, we proposed the RSRI model for patent retrieval fusion, which has lower dependence on the single retrieval engine. The RSRI model and the results merging strategies can complete the retrieval task effectively, obviously simplify the retrieval fusion process and enhance the retrieval efficiency. Experiments on a collection of patent documents showed that our RSRI model improved the quality of patent retrieval system.

REFERENCES

Belkin, N.J., P. Kantor, E.A. Fox and J.A. Shaw, 1995. Combining the evidence of multiple query representations for information retrieval. Inform. Process. Manage. Int. J., 31: 431-448.

Callan, J.P., Z.H. Lu and W.B. Croft, 1995. Searching distributed collections with inference networks. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 09-13, ACM, Seattle, Washington, pp: 21-28.

Fox, E.A. and J.A. Shaw, 1999. Combination of multiple searches. Proceedings of the 2nd Text Retrieval Conference (TREC-2), (TRC'99), National Instituted of Standards and Technology Special Publication, pp: 151-173.

Hu, X., 2006. Research on Information Retrieval Based on Language Model and Reranking for Retrieval Results. Harbin Institute of Technology, Harbin.

Jiang, Y., J. Lei and G. Zhou, 1999. Technological Innovation Management. Enterprise Management Press, Beijing.

Jin, B. and H.F. Teng, 2007. Case-based evolutionary design approach for satellite module layout. J. Sci. Ind. Res., 66: 989-994.

Jin, B., H.F. Teng, Y.J. Shi and F.Z. Qu, 2007. Chinese patent mining based on sememe statistics and key-phrase extraction. Adv. Data Min. Appl., 4632: 516-523.

Jin, B., H.F. Teng, Y.S. Wang and F.Z. Qu, 2008. Product design reuse with parts libraries and an engineering semantic web for small- and medium-sized manufacturing enterprises. Int. J. Adv. Manuf. Technol., 38: 1075-1084.

Jin, B., Y.J. Shi and H.F. Teng, 2010. Textual chunk retrieval and packing in pattern. Proceedings of the 2nd IEEE International Conference on Information Management and Engineering, May 2-4, Chongqing, China, pp: 301-305.

Liu, Y., S. Liu and B. Yu, 2008. Chinese patent search model and experiment for examination task. Appl. Res. Comput., 25: 1483-1484.

Rasolofo, Y., F. Abbaci and J. Savoy, 2001. Approaches to collection selection and results merging for distributed information retrieval. Proceedings of the 10th International Conference on Information and Knowledge Management, Oct. 05-10, ACM, New York, USA., pp: 191-198.

Xu, J. and W.B. Croft, 1999. Cluster-based language models for distributed retrieval. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, ACM, New York, USA., pp: 254-261.

Xu, Y., L.C. Hu, W. Zeng and B. Jin, 2010. Web-service-based parametric design reuse for parts. Int. J. Adv. Manuf. Technol., 46: 423-429.