Intelligent Retrieval System for Patent Analysis - the Combined use of Bibliographic Coupling and Text Mining

Su-Houn Liu [1] **Hsiu-Li Liao** [2] **Jing-Wen Hu** [1]

[1]Chung Yuan Christian University, Taiwan
[2]Yuan Pei University, Taiwan

Abstract

When technology developments accelerate in recent year, automatic tools for assisting patent engineers or decision makers in patent analysis are in great demand. By composing both bibliographic coupling and text mining approaches, this study proposes an intelligent patent retrieval system that considers a hybrid structure for higher search accuracy. An experimental prototype called HPRS (Hybrid Patent Retrieval System) was developed. Testing indicates that the HPRS has significantly increased the accuracy of patent retrieval compared to traditional patent search methods. We believed that our works have provided a feasible architecture for an intelligent patent retrieval system.

Keywords: patent search, text mining, bibliographic coupling

1. Introduction

With an ever increasing pace of technology development, patent search and analysis has become a prominent task in both legal and innovation management [6]. Unfortunately, it remains time consuming for inventors or searchers to find out the right patents that they really want. To improve the search process, various intelligent data and text mining tools applied to patent analysis have been around for quite a while now [12]. However, since mining tools always try to analyze content using mathematical methodologies, they overlook the fact that patent records are combinations of both structured and non-structured data [8]. The researchers conducting this study believed that by integrated bibliographic coupling mechanism to structured data [9] with text mining tools on analyzing non-structured data, it is possible to construct an enhanced patent retrieval system for use in patent analysis.

This study developed an experimental prototype called HPRS (Hybrid Patent Retrieval System). To test the effectiveness of HPRS prototypes, this study conducted a series of tests with various system settings and parameters. Research implications and issues of future works are discussed in the last section.

2. HPRS Prototype

The core of the HPRS is a combination of its field matching engine and text mining engine by a weighting model. By entering several origin patent records, HPRS will scans the target patent records using Pipelines on its two engines. The Weighting model then combines the similarity rankings generated by the Pipelines to come out the final similarity ranking. Patents most similar to the origin patents are rec-

Proceedings of the 11th Joint Conference on Information Sciences (2008)
Published by Atlantis Press

ommended to searchers via the presentation layer. Fig. 1 illustrates the architecture of the HPRS.

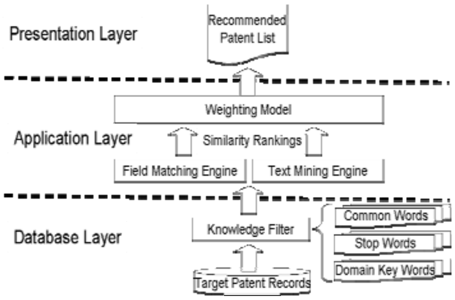

Fig. 1: HPRS (Hybrid Patent Retrieval System) architecture

2.1. Field Matching Engine

For processing the structured data in the patent database, the Field Matching Engine comprises seven Pipelines. They are the Inventor Pipeline, Assignee Pipeline, Examiner Pipeline, IPC Pipeline, UPC Pipeline, Forward Citation Pipeline, and Backward Citation Pipeline. The Field Matching Engine uses a bibliographic pattern discovering algorithm to identify clusters of related patent records in a collection [1][9]. First, from the origin patents, the instances set of specific data field was treated as thought it was the citation list in a document. For each Pipeline, the similarity of the pseudo-document to each patent in the target database was calculated using the following familiar measure:

$$S_{ot} = \frac{\left|I_o \cap I_t\right|}{\left|I_o\right| + \left|I_t\right| - \left|I_o \cap I_t\right|} \qquad (1)$$

Where I_o denotes the set of instances "cited" by the pseudo-document "$_o$" and I_t denotes the set of instances of target patent " $_t$" in that specific data field. For example, on the Inventor Pipeline, if we have five inventors from the origin patents ($I_o=5$), and the target patent have three inventors ($I_t=3$), if 2 inventors of the target patent also exist in the inventor set of the origin patents, then $I_o \cap I_t = 2$. Therefore the similarity S_{ot} should be 33.33%.

2.2. Text Mining Engine

Regarding patent analysis, text mining is used as a data-processing and information-extracting tool [10][12][15]. The Text Mining Engine has five Pipelines: Title Pipeline, Abstract Pipeline, Patent Claim Pipeline, Patent Claim1 Pipeline and Detail Description Pipeline. Since the definition of similarity may be quite different between different patent searchers [13] [14], the HPRS allows users to select which Pipelines to be enabled in the text mining analysis. To further integrate user expertise into the analysis, the HPRS also allows users to enter self-assigned keywords or disable certain keywords during their search process. The HPRS Text Mining Engine is implemented using the vector space model component of MagaPuter^{TM}. The core of the engine is adaptation to text categorization based on the formula of Rocchio [11].

2.3. Weighting Model

The weighting model is responsible for combining the results from the Field Matching and Data Mining engines. However, it is extremely difficult to make direct comparisons between results generated by different Pipelines. To avoid this problem, the results generated by both the Field Matching and Data Mining engines are converted to McCall T scores [5] by the Weighting Model. This conversion sets the means of the distributions to 50 and the standard deviations to ten. Additionally, this conversion corrects the skewness of the distributions, making them normal in shape. The final result was then obtained by the calculation of the confidence index (CI) of each patent record. After each selected Pipeline produced a normalized T score for the similarity degree of each target patent, the final score of similarity (FSS) was calculated using the weighting parameters entered by the searcher before initiating the process. After the calculation of FSS for each target patent, the Weighting Model sorts these patents according to their FSS, and presents the results via the Presentation Layer.

3. Testing of the HPRS Prototype

We prepare a set of target patent records to verify the effectiveness of the HPRS in patent retrieval. Despite the existence of several benchmark collections on patent records (such as WIPO-Alpha), it is necessary to obtain an independent source of test data because bibliographic data in an ordinary patent document are presented in the form of text fields rather than structured data fields. For example, in the Assignee field, “IBM”, “International Business Machines” and “International Business Machines Corp.” indicate the same company. For the Field Matching Engine to function properly, the bibliographic data in the target patent records must be pre-processed prior to the analysis. In this study, we have the privilege to access on a patent records set that was prepared by a large research institute as part of a government funded project on constructing a patent map related to the world’s GMO (genetically modified organism) development. Although 100% clarification of all bibliographic data in every patent document is virtually impossible (for example, it is difficult to know whether “J. H. Lin” and “Jin-Houng Lin” are the same inventor), the 10 experts in the research team did their best to generate an accurate patent map. Their efforts made their patent record set ideal for testing the HPRS prototype.

Patent documents are extracted from the U.S. Patent and Trademark Office (USPTO: www.uspto.gov) database. In all, 267,156 patents were collected with the reference period from 1976/1/1 to 2002/8/20. Among those 267156 patents, 408 patents were further selected by these experts as the core patents. Those 408 patents were then classified into 23 technology classes and 11 effect classes. Nine technology classes that have more than 30 patents in it were selected by this study to be our test data set.

For patents on each classes, our study randomly selects 1/3 as the training data set. For these nine technology classes, 5 classes have 10 patents, 3 classes have 20 patents and only 1 group (FN5) has 40 patents as the training data set. The other 2/3 was used as the target data set. After we input those training data set into the HPRS as the origin patents, we expect that the HPRS can successful recall the rest 2/3 patents in the target data set from the total 267,156 patents.

4. Experiment Process

In order to verify the effectiveness of our hybrid HPRS prototype, we conduct the test procedure as following:

Stage 1: Establish the comparison baseline

On the HPRS Data Mining Engine, there are five text fields that can be selected to perform text mining. They can form 29 meaningful combinations. We conduct testing on all these 29 combinations to identify which combination can generate the best result. The combination and the result will be used as the baseline to identify the effectiveness of the HPRS which integrate bibliographic coupling with the data mining.

Stage 2: Identify the effectiveness of each bibliographic data field

On the HPRS Data Matching Engine, there are seven bibliographic data field (Pipelines) can be selected to perform data matching. On this stage, we try to identify the effectiveness of each Pipeline on retrieve related GMO patents.

Stage 3: Experiment on the hybrid model

On this stage, we will test the effectiveness of using pipeline combinations on retrieve patents on the target data set. We then compare the result with the result that generate by stage 2 to see whether the combination approach performed better or not.

Stage 4: Experiment on different weighting parameters

HPRS allowed searchers to set different weighting for each Pipeline. On those previous stages, we set all pipelines to have the same weighting. But on stage 4, we will try different weighting parameters to see whether it will generate a better result.

5. Findings

Stage 1: Establish the comparison baseline

On our previous study [3][4], the text mining engine on the HPRS is an effective tool to do automatic patent classification on those GMO patents. But under a small number of training data set, pure text mining approach can not perform very well. On the 29 meaningful text field combinations, one of the best performers is the T+D (Title Pipeline and Detail Description Pipeline) combinations. But, as indicated on fig. 2, when it recommends 15,000 patents, the recall rate is only 40%.

Stage 2: Identify the effectiveness of each bibliographic data fields

On stage 2, we turn on the Pipelines of the Data Matching engine one at a time to filter the patents recommend by text mining engine. Some of the Pipelines (Backward Citation, Assignee and Examiner) seem not be able to improve the effectiveness of the data mining engine (compared to the result on stage 1). But the other fields (UPC, IPC, Inventor, and Forward Citation) can improve the number of recommendation significantly (Table 1). This result indicated that at least some bibliographic data fields can be used to improve the effectiveness of pure text mining results.

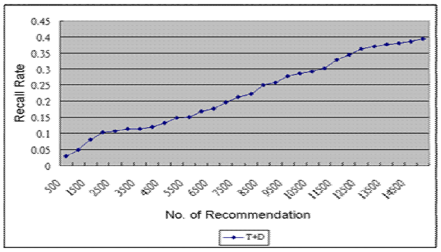

Fig. 2: Recall rate for T+D combination

Tab. 1: Improvement on each bibliographic data fields

Recall Rate = 50%			Recall Rate = 70%			Recall Rate = 90%		
Combina-tion	Improve-ment	Std. Div	Combina-tion	Improve-ment	Std. Div	Combina-tion	Improve-ment	Std. Div
UPC	-8087	4140	UPC	-14108	9613	UPC	-34679	25248
IPC	-3937	3367	IPC	-7519	8617	IPC	-19010	14790
inventor	-3242	4714	F.Citation	-5204	8797	F.Citation	-9423	16346
F.Citation	3093	4327	Inventor	-4343	9540	Inventor	-7670	15586
Assignee	-1504	8573	B.Citation	-2547	8052	B.Citation	-4829	14625
B.Citation	-1376	3443	Assignee	-997	10303	Assignee	-1780	15628
Examiner	2338	15035	Examiner	87081	76787	Examiner	122432	28929

Stage 3: Experiment on the hybrid model

On stage 3, we try to further improve the effectiveness of the HPRS prototypes by combining several bibliographic Pipelines. Three Pipelines (Backward Citation, Assignee and Examiner) are excluded due to their low improvement rate on stage 2. The rest four Pipelines can combine into 11 combinations and the result is listed on Table 2. Compare to the result of stage 2, the result of stage 3 indicates further improvement on number of recommended patents to reach certain recall rate. The standard deviations were also improved by nearly 50% which indicates that by combine several bibliographic Pipelines, the HPRS prototypes can generate more accurate and more stable results. When the recall rate was set on 70% and 90%, the I+UPC+F (Inventor Pipeline, UPC Pipeline and Forward Citation Pipeline) combination can generate best result compare to other combinations. On 50% recall rate, I+IPC+UPC+F(Inventor Pipeline, IPC Pipeline, UPC Pipeline and Forward Citation Pipeline) can generate better result than the I+UPC+F combination. But the difference on improvement rate was not impressive and the Standard Deviation is almost the same.

Tab. 2: Improvement on combinations of bibliographic data fields

Recall Rate = 50%			Recall Rate = 70%			Recall Rate = 90%		
Combina-tion	Improve-ment	Std. Div	Combina-tion	Improve-ment	Std. Div	Combina-tion	Improve-ment	Std. Div
I+IPC +UPC+F	-4855	2359	I+UPC+F	-7625	3635	I+UPC+F	-18636	15066
I+UPC+F	-4459	2357	I+IPC +UPC+F	-7470	3783	I+IPC +UPC+F	-17998	12796
IPC+UPC +F	-3573	2475	IPC+UPC +F	-5502	3619	UPC+F	-13632	11249
I+IPC +UPC	-3551	2325	I+IPC +UPC	-5344	3831	IPC+UPC +F	-13119	10852
I+IPC+F	-3481	2419	UPC+F	-4888	3235	I+UPC	-12573	12215
UPC+F	-3054	2297	I+UPC	-4289	3733	I+IPC +UPC	-12075	10885
I+UPC	-2620	1745	I+IPC+F	-3894	3194	IPC+UPC	-6453	7979
I+F	-1536	1878	IPC+UPC	-2518	3568	I+IPC+F	-6222	10845
IPC+UPC	-1451	2221	I+F	-2076	2251	IPC+F	-3856	7731
IPC+F	-1392	2038	IPC+F	-1805	4129	I+F	-1861	5585
I+IPC	-915	1714	I+IPC	-1737	3808	I+IPC	-1558	7670

Stage 4: Experiment on different weighting parameters

When the searcher is conducting a patent search, his expertise may indicate that on this specific search, some bibliographic data fields may be more important than other fields. In order to utilize the searchers' expertise, HPRS prototype allow the patent searcher to set weighting parameter for each bibliographic field selected. The HPRS prototype will calculate the result on a comparative basis. On stage 4, we try to set different combination of weighting parameter and compare their result to the result of stage 3. As indicated on fig. 3, none of the combinations on stage 4 can out perform the best result of stage 3. This result may indicate that unless the searcher has a strong evident that convince him to set uneven weighting parameters, to set all fields on same weighting may be a safer way of doing his patent search on HPRS.

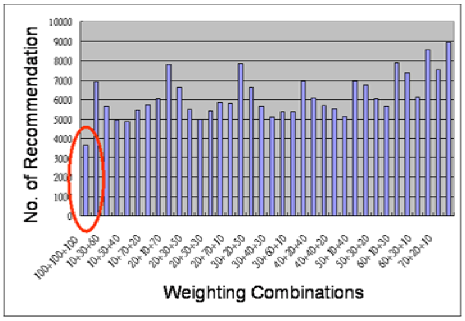

Fig. 3: Comparison of the result between Stage 3 and Stage 4.

6. Conclusions and Future Work

After we integrated data matching on single bibliographic data field on stage 2, under the same recall rate, the HPRS prototype were able to substantially reduce the number of patent recommended. On stage 3, by introduce different combinations of the bibliographic data fields we further increased its accuracy. On some best combination, the HPRS had reduced the number of patent recommended on stage 2 by 90%. The result of stage 4 reveals that on our GMO data source, an even weighting parameter setting can perform better than any other weighting parameter setting. Even though the testing result may not stand for patents from other technology domain, but we believed that our study had proved that the hybrid approach is a feasible architecture in patent retrieval applications.

Overall, the results show that hybrid approach may help to reduce the necessary effort of patent retrieval. Furthermore, the utility of the proposed approach can be extended and/or elaborated far beyond the scope of patent search. It can also be used to overcome the drawbacks of using text mining technique on searching other semi-structured documents, such as journal papers. Another promising themes of further research is on how to (or if it is possible to) train technology or patent experts to use the HPRS prototype in order to speed up their searching process. When an expert work collaboratively with the HPRS, his expertise can be transform into setting of keywords, weight parameters, data fields selection and selection of a proper training data set. Even though further testing is required, we believed that the expert involvement is possible and can help to improve the effectiveness of HPRS significantly.

7. References

[1] Bichteler,J. Eaton, E.A., "The Combined Use of Bibliographic Coupling and Cocitation for Document Retrieval," Journal of the American Society for Information Science, 31(4), pp. 278-282, 1980.

[2] Chemical Week, "Intellectual Property Software: Anything Extra with That ?," Chemical Week, 23, 2002.

[3] Chih-Chiang Kao, "A Study of Combining Automatic Document Classification-Example on Patent Document," Master Thesis, Cung-Yuan Univ., 2004.

[4] Chun-Hsiang Lee, "A Study of Applying Data Mining Classification Techniques to Patent Analysis," Master Thesis, Chung-Yuan Univ., 2003.

[5] Edwards, W., "The theory of decision making," Psychological Bulletin, 51, pp. 380–384, 1954.

[6] Germeraad, P. B., and Lorraine Morrison., "How Avery Dennison Manages Its Intellectual Assets," Research・Technology Management, pp. 36-43, 1998.

[7] Homan H.S., "Making the case for patent searchers," Searcher, pp. 8-14, 2004.

[8] IBM Corp., Intelligent Miner for Text: Getting Started, 1998.

[9] Kessler, M.M., "Bibliographic Coupling Between Scientific Papers," American Documentation, 14, pp.10-25, 1963.

[10] Lee, S., Lee, S., Seol, H. & Park, Y., "Using patent information for designing new product and technology," R & D Management, 38(2), pp. 169-181, 2008.

[11] Rocchio, J.J., "Relevance feedback in information retrieval," in Salton, G. (Ed.), The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, Englewood Cliffs, NJ, pp. 313-23, 1971.

[12] Tseng, Y., Lin, C. & Lin, Y, "Text mining techniques for patent analysis," Information Processing & Management, 43(5), pp.12-16, 2007.

[13] Tzeras K, Hartmann S.: Automatic indexing based on Bayesian inference networks. In Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval, pp. 22-34, 1993.

[14] Yang Y., "Noise reduction in a statistical approach to text categorization," In Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, pp. 256–263, 1995.

[15] Yoon, B., Phaal, R. & Probert, D., "Morphology analysis for technology roadmapping: application of text mining" R & D Management, 38(1), pp. 51-64, 2008.