Development of a Patent Matching System Using a Hybrid Approach

Su-Houn Liu Hsiu-Li Liao Chou-Chih Hsieh[1]

[1]Chung Yuan Christian University
Chung Yuan Christian University
LearningTech Corp

Abstract

There were many researches about applying various data mining or text mining tools to patent analysis, and there were many scholars and experts have verified the accuracy and the feasibility of those tools. However, since mining tools always tried to analyze the content using some mathematic methodology, such as linguistic algorithms, they neglect the fact that patent records are combinations of both structured and non-structured data; it contains not only the non-structured descriptive text but also many structured data related to each patent, such as inventors, assignees and citation information… etc. In another word, mining methodology tent to neglect this import features of patent records and handled them as pure text.

This paper proposes a hybrid approach to conduct patent matching process. In this study, an experimental prototype call PMS (Patent Matching System) was developed by composing both data matching and mining approach. By entering several origin patents, the PMS will scan the patent database to generate a similarity ranking table, and then patents that most similar to those origin patents will be suggested to the user. As our sample testing reveals, the PMS achieved a remarkable patent matching capability, and show potential for further improvement.

Keywords: patent matching, data mining, patent analysis

1. INTRODUCTION

There are evidences that patent has become very important given by increasing lawsuits of patent. Accordingly, patent has become the critical weapon on the war of knowledge-based competition [5]. Unfortunately, it was a time-consuming effort for patent searchers to find out the patents that he really wanted; it was not only because of the mass quantity of patent records in the database, but also because it required the searcher's specialty and experience to reduce the range of necessary patents by adopting various searching methods step by step [7]. After the patent search, the patent searcher has to read through all the patents in the search result in order to exclude those un-related patents, and it is really a heavy job to read patents piece by piece.

There were many researches about applying various data mining and text mining tools to patent analysis, and there were many scholars and experts have verified the accuracy and the feasibility of those tools [2][3]. However, since mining tools always tried to analyze the content using some mathematic methodology, such as linguistic algorithms, they neglect the fact that patent records are combinations of both structured and non-structured data; it contains not only the non-structured descriptive text but also many structured data related to each patent, such as inventors, assignees and citation information… etc. In another word, mining methodology tent to neglect this import features of patent records and handled them as pure text [4][6].

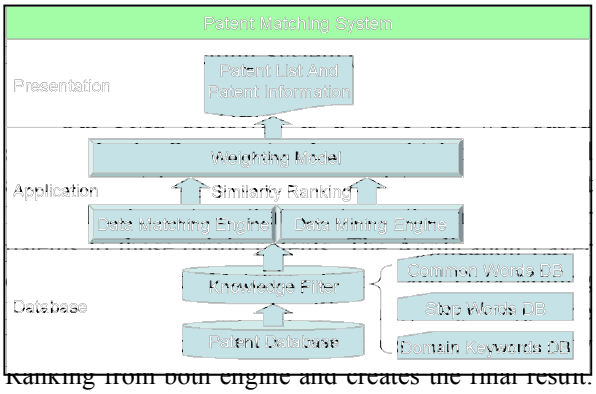
In order to analyze patents, one should remember that patents are both structured and non-structured content of data. The researchers of this study thus believed that by integrated data matching on structured data with text mining tools on analyzing non-structured data, we may construct a better patent matching tool to use on patent analysis process.

The purpose of this paper is to propose a hybrid approach to conduct patent matching process. In this study we have build up a prototype system according to the architecture that we have proposed. A patent matching testing was conducted with sample patent database to verifying the effectiveness of this PMS prototype.

2. PMS Prototype

The core of our Patent Matching System (PMS) is the combination of the data matching engine and data mining engine by its weighting model. By entering several origin patents, the PMS will scan the patent database to generate a similarity ranking table,

and patents that most similar to those origin patents will be presented to the user. Figure 1 shows the architecture of the PMS.



Ranking from both engine and creates the final result. The Database Layer provides the patent raw data with the help of Knowledge Filter. The Knowledge Filter contains Common Word DB, Stop Word DB and Domain Keywords DB. These DB can be loaded into the Knowledge Filter and eliminate or emphasize certain words to increase the accuracy of our matching and mining engines.

2.1. Presentation Layer:

In order to present the similar patents to the user, the PMS interface (Figure2) is separated into three main parts; one shows the recommended patent list while the other shows the detail information of certain patent in the list that was selected by the user. Besides, the interface also shows the keywords that were used by the PMS's mining engine to construct its similarity ranking. User can refresh the result by alter these keywords list.

2.2. Application Layer:

- Examiner pipelines: Count for the same USPTO examiners between the origin patents and suggested patents.
- Forward Citation pipelines: Count for the similarity of the forward citations between origin patents and suggested patents.
- Backward Citation pipelines: Count for the similarity of the backward citations between origin patents and suggested patents.
- IPC pipelines: Count for the same IPC classifications between origin patents and suggested patents.
- UPC pipelines: Count for the same UPC classifications between origin patents and suggested patents.

B. Data Mining Engine:

For processing those non-structured data in the patent database, our mining engine will generate a semantic network map base on the selected keywords after analyzing those origin patents entered by the user [10][11][12]. This Data Mining Engine is implemented by using the vector space model component of MagaPuter™. Since the definition of similar patent may be quite different between different users, the PMS allow its users to choose which patent fields (Title, Abstract, Patent Claim or Detail Description) will be input into the text mining analysis. To integrate the user's expertise into the analysis, the PMS also allow its user to enter self-assigned keywords.

C. Weighting Model:

Weighting model is responsible for combining the derived results from both Data Matching Engine and Data Mining Engine. The final result was creating by the calculation of the confidence factor of each patent record [9]. When both matching and mining engines came out a similar ranking, the ranking was accepted by the Weighting Model as the final result. When the ranking are quite different, the final ranking was calculated by the weighting parameters that entered by user before the analysis. After the calculation of new ranking, the Weighting Model recommends a list for those most similar patents through the Presentation Layer.

2.3. Database Layer:

The PMS's Database Layer consists of two main components, including:

A. Patents Database:

The Patent Database is responsible for storing patent data for further analysis. Since our prototype are developed to process only US patents, only patents on USPTO format can be processed by the PMS.

B. Knowledge Filter:

The Knowledge Filter of the PMS has three databases:

- Stop words DB: This DB is responsible for storing stop words. When it was turn on, The filter will decrease the noise of matching result by eliminate those un-meaningful stop words.
- Common words DB: This DB is responsible for storing common words that usually used in patent documents. When it was selected, the Filter will treat these common words as stop words.
- Domain keywords DB: This DB is responsible for storing domain keywords that represented critical meanings in certain technology domain. When it was selected, the Filter will increase the weight of those keywords on the calculation of the similarity ranking.

3. Evaluations

3.1. Test Setting

With PMS prototype, we prepare a set of testing data to verifying the PMS' effectiveness on patent matching. Our testing data contained 4148 US patents related to the semiconductors manufacturing process. We invite two experts from TSMC to assist our study. The first expert was asked to set the weighting parameters on the PMS. Table 1 shows the experts' suggestion.

Pipeline	Weight
Examiner pipeline	10
Inventors pipeline	20
Assignees pipeline	20
IPC pipeline	50
UPC pipeline	50
Forward citation pipeline	150
Backward citation pipeline	100
Data mining engine	200

Tab. 1: the weighting parameters on the PMS

In the data mining engine we also select all patent fields and set given experts' suggestion about their weight on each fields. The expert also provides several self-assigned keywords. They are "lead frame", "leadframe" and "non-leaded".

Patent field	Weight
Bibliography data	10
Title	20
Abstract	20
Detail Description	20
Claims	30
First claim	50

Tab. 2: the weighting parameters on the Data Mining Engine

The other experts provided 8 patents as our origin patents. Those patents' patent numbers are 06683368, 06661087, 06580165, 06507120, 06307256, 06294838, 06060769, and 06683368. And the expected matching result contains 30 patents that the expert identifies them as close related to those 8 origin patents. Those patents' patent number are 06710454, 06707136, 06700187, 06689640, 06683375, 06664615, 06657288, 06646316, 06630733, 06597059, 06583035, 06570251, 06566168, 06544817, 06518650, 06504236, 06495908, 06448110, 06329710, 06321976, 06310388, 06255720, 06229204, 06215177, 06204163, 06184574, 06072228, 06008531, 05907184, and 05717246.

To verifying the PMS's effectiveness, we defined three effective indexes as fallow:

- Recall ratio: Correct suggested patents / expected matching results
- Precision ratio: Correct suggested patents / total suggested patents

- Reading ratio: total suggested patents / Total number of patents in patent Database

3.2. Results

The suggestions of the PMS create quite remarkable results show on Table 3:

No of Patents Suggested	Correct Suggested Patents	Recall Ratio	Precision Ratio	Reading Ratio
30	26	0.867	0.867	0.007
60	29	0.967	0.483	0.014

Tab. 3: The result of the PMS effectiveness evaluation

When suggested patents numbers was set to 30 patents, the recall ratio is 0.867. It means that almost 86.7% expected patents come out from the suggested matching patents. When the max suggested patents go up to 60 patents, we can find that the Precision ratio decrease while recall ratio increase. The recall ratio rises to 0.967. Under this scenario, the reading ratio is 0.014. It means that readers can just read 1.4% of suggested patents and covered almost 96.7% of expected patents.

4. Conclusion

The results show that our PMS prototype can help to reduce the necessary effort of the patent searchers significantly. Nevertheless, PMS is still improvable. During this study, we never discuss the multiple combinations of pipelines and patent fields. And there may be better data mining algorithms or tools that can be embedded into our prototype. Our testing shows that PMS can achieve a remarkable patent matching capability. Even though the testing result shows in this article may not stand for patents from other technology domain, but we believed that our study have proving a feasible architecture in matching patents. And in this case, we can draw the conclusion that PMS can be seen as an effective tool for patent searchers to manipulate massive patent records.

5. References

[1] Chemical Week (2002). Intellectual Property Software: Anything Extra with That?, Chemical Week, 23.

[2] Chih-Chiang Kao, *A Study of Combining Automatic Document Classification-Example on Patent Document*, 2004.

[3] Chun-Hsiang Lee, *A Study of Applying Data Mining Classification Techniques to Patent Analysis*, 2003.

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

[5] Japan Patent Office (2000). Patent Management in Enterprises. Asia-Pacific Industrial Property Center.

[6] Fattori, M., Giorgio Pedrazzi, and Roberta Turra (2003). Text mining applied to patent mapping: a practical business case, World Patent Information, 25, 335-342.

[7] Germeraad, P. B., and Lorraine Morrison. (1998). How Avery Dennison Manages Its Intellectual Assets, Research • Technology Management, 36-43.

[8] Li-Ping Jing, Hou-Kuan Huang, Hong-Bo Shi., "Improved Feature Selection Approach TFIDF In Text Mining," *Proceedings of the First International Conference on Machine Learning and Cybernetics*, November 2002.

[9] 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*, 1993, pp. 22-34.

[10] 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*, 1995, pp. 256–263.

[11] Yang Y., Pedersen J.O., "A comparative study on fetue election in text categorization," *In proceedings of ICML-97, 14th International Conference on Machine Learning*, 1997, pp. 412-420.

[12] Yang Y., Slattery S., Ghani R., "A study of approaches to hypertext categorization," *J. In-tell. Inform. System*, 18, 2/3, 2002, pp. 219-241.