Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Development of a Multilingual Text Mining Approach for Knowledge Discovery in Patents

Chung-Hong Lee
Department of Electrical Engineering,
National Kaohsiung University of Applied Sciences,
Kaohsiung, Taiwan
leechung@mail.ee.kuas.edu.tw

Hsin-Chang Yang
Department of Information Management,
National University of Kaohsiung,
Kaohsiung,
Taiwan
yanghc@nuk.edu.tw

Yi-Ju Li
Department of Electrical Engineering,
National Kaohsiung University of Applied Sciences,
Kaohsiung, Taiwan
leeyiju@dml.ee.kuas.edu.tw

***Abstract*—In this paper we describe our work on developing a novel technique for discovery of implicit knowledge about patents from multilingual patent information sources. In this work we developed a system platform to support locating similar and relevant multilingual patent documents. The platform was implemented using a multilingual vector space based on the latent semantic indexing (LSI) model, and utilizing collected professional Chinese-English parallel corpora for training the system model. These multilingual patent documents could then be mapped into the semantic vector space for evaluating their similarity by means of text clustering techniques. The preliminary results show that our platform framework has potential for retrieval and relatedness evaluation of multilingual patent documents.**

***Keywords*—Text mining, Multilingual patent retrieval, Patent retrieval, Latent semantic indexing, Document clustering**

I. INTRODUCTION

For many organizations, patent information is a crticle source of technical information. Analysis of patent information has been regarded as a key process to discover the technical knowhow for product development. As a result, some companies have applied patent analysis to understand the needs of the market and technology trend. Yeap [34] addressed the issue that patent analysis can be used to gain strategic advantages, and help managers assess development priorities among possible R&D project proposals. The importance of patent analysis in strategic planning has also become increasingly apparent. European Patent Office (EPO) disclosed that "patents reveal solutions to technical problems, and they represent an inexhaustible source of information: more than 80 percent of man's technical knowledge is described in patent literature"[6]. Figure 1 illustrates a generic model of patent analysis process. The process normally started with the step of manually determining the subject of patents. Then analysts can search relevant patent information using patent information systems to carry out patent analysis.

In addition, inventors of internationally distributed information networks need tools and methods that will enable them to discover, retrieve and understand relevant patent information, in whatever language. Traditional techniques of multi- and cross-language patent retrieval are mostly based on the use of dictionaries and translation techniques. Unfortunately, due to the language difficulties, it is normally hard to quickly find related patents produced from other countries in a stand-alone patent information system. In addition, patent documents are usually domain specific and full of technical terminologies, and existing dictionaries are insufficient to process these highly specialized and massive technical terms. The view is taken, therefore in this work, we establish a novel system platform using text mining methods to support locating similar and relevant multilingual patent documents. By means of a developed unified vector space based on the latent semantic indexing (LSI) model for multilingual document projection, we are able to handle the retrieval and discovery of more similar and relevant patent documents in various languages. These multilingual patent documents can then be mapped into the semantic vector space for evaluating their similarity by text clustering techniques.

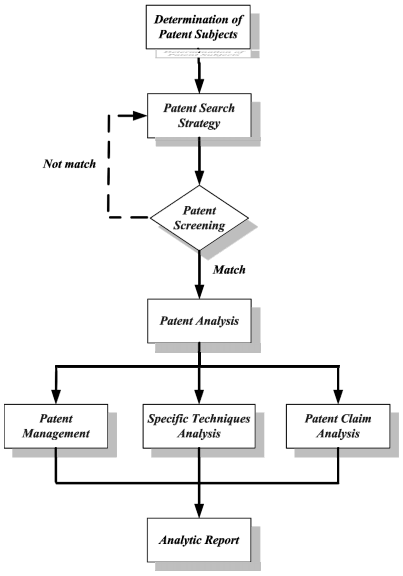

Figure 1. Patent anlysis process

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

The platform includes synonym processing functions to improve its accuracy. In addition, we have collected several professional Chinese-English parallel corpora (i.e. *EE Times Asia*, a professional document collection in the field of Electrical and Electronics Engineering) for training the system model. Also, we utilized International Patent Classification (IPC) codes to evaluate two clustering techniques in our platform. The rest of the paper is organized as follows: Session 2 reviews related work in the domain. Session 3 presents our system model. Session 4 shows our experimental results, and the last section presents our conclusion.

II. Attempted Solution- A Multilingual Patent Text-Mining Approach

In this section, we describe our developed approach and system platform for text mining and retrieval of multilingual patents. Figure 2 illustrates our system framework, including three main phases: namely the construction of multilingual semantic space, process of patent document mapping and text clustering modules.

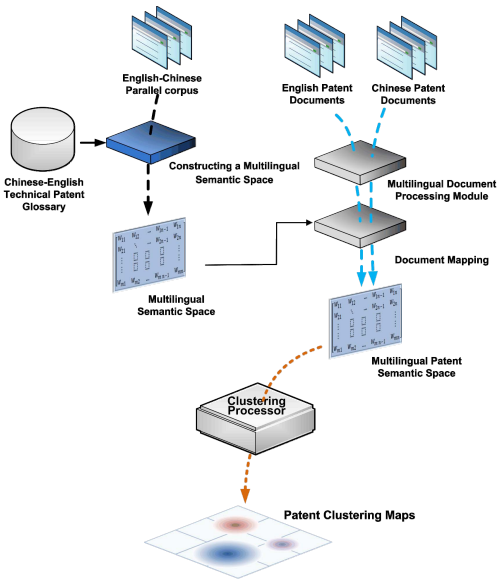

Figure 2. System framework

For document preprocessing, the programs of CKIP (Chinese Knowledge and Information Processing) and CLAWS (Constituent Likelihood Automatic Word-tagging System) are used to add proper part-of-speech information to each word for easily indexing collected documents. Subsequently, nouns and noun phrases are extracted from each document for indexing work. In the step of synonym parsing, we employed the Chinese-English Technical Patent Glossary to aggregate synonym terms in the same class. The vector space model generates a term-to-document feature-appearance matrix where rows are the features and columns are document vectors, and the values of each matrix elements are weighted by TF (Term Frequency). Finally, we utilized LSI to decompose the term-document matrix and construct a multilingual semantic space onto the K dimensions space.

In the patent document mapping phase, initially, we utilized IPC codes to classifying patent documents manually. In multilingual document processing module, the major content text of each patent (mainly the abstract and claim) were retained from patent documents for further analysis. Then nouns and noun phrases were extracted from each abstract and claim for encoding documents into document vectors. Each patent-document vector can be projected into an unified semantic space for multilingual text mining.

In order to further allow users to find the ranking of candidate patents for knowledge discovery about a new invention, we used two clustering techniques, including hierarchical agglomerative clustering (HA) and self-organizing maps (SOM) methods, to perform patent text mining. The performance of these two clustering techniques would be evaluated by the measures of recall and precision tests.

III. Experimental Results

In this work, we used 2,000 documents (1,000 Chinese texts and 1,000 English ones) in Chinese-English parallel corpora collected from *EE Times Asia* [1], a professional document collection in the field of Electrical and Electronics Engineering, to construct a multilingual semantic space. We also collected 1,200 patent documents (600 Chinese texts and 600 English ones) from Intellectual Property Office, R.O.C. (TIPO) and United States Patent and Trademark Office (USPTO) as our test corpora. Furthermore, all patents have been manually assigned with cluster information according their IPC codes.

To investigate the effects of developed synonym processing mechanism, we compared the clustering results of our proposed system, in terms of whether or not adding the process of synonym elimination.

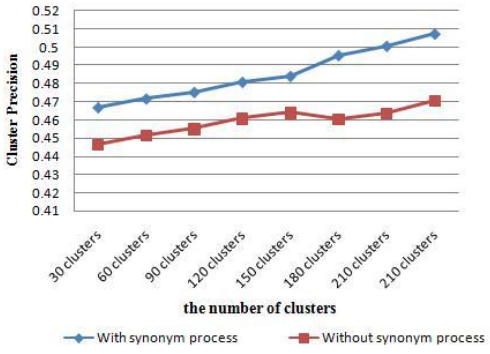

Figure 3. Results of precision measures of developed clustering techniques

As shown in Figure 3, it is obvious that the precision performance achieved by the experiments using synonym processing techniques is superior to that of clustering system without utilizing any synonym processing methods.

[1] http://www.eetimes.com

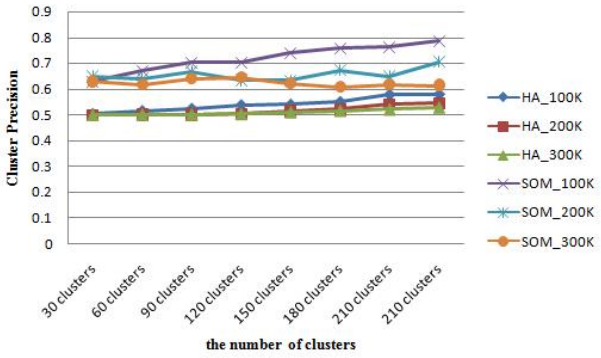

Figure 4. Results of precision measures of developed clustering techniques

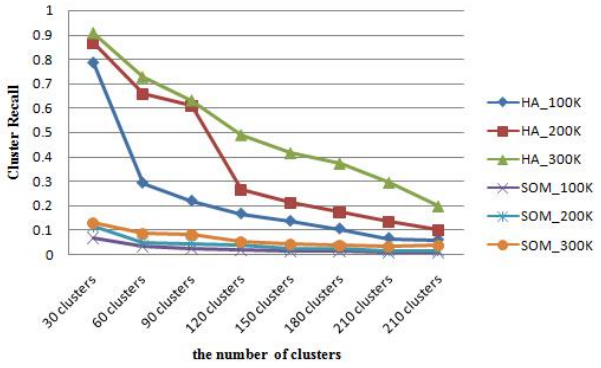

Figure 5. Results of recall measures of developed clustering techniques

Figure **6.** Results of F-measures of developed clustering techniques

Recall and precision are both considered for determining the value of K parameter. As shown in Figure 4-5, hierarchical clustering method has a better performance while K is set to 200, and the one of SOM is set to 300. The experimental results of F-measures are shown in Figure 6, and the clustering technique based on hierarchical clustering (HA_200k) outperforms the SOM based approach (SOM_300k).

The results of experiments have demonstrated that our system can provide a way for fully automated functions of cross-language (Chinese and English) patent retrieval, in which no translation of patent documents was required. The system allows users use to find related patent documents in either Chinese or English language. According to the experimental results (shown in Figure 4-6), in this work the experimented clustering technique based on hierarchical clustering outperformed the SOM based approach, and it required comparatively fewer dimensions to carry out the task.

IV. Related Work

Most of the related research work on patent information processing has been focused on developing patent document retrieval systems to supporting patent analysis. In the late 1970s with advances in the earlier text retrieval models, the field of *intelligent information retrieval* was investigated as a major empirical method to patent information processing. For instance, Attar and Fraenkel [1] used local feedback techniques to find relevant patents, and Osborn [25] utilized the SMART system, based on natural language processing techniques, to perform information search of patent documents. In order to avoid prohibitively expensive search time and improve accuracy for a large collection of patent documents, Larkey [19] divided a whole collection of patent documents into several sub-collections according to their IPC codes. Furthermore, Ryley proposed latent semantic indexing (LSI) [3] methods combined with the vector space model (VSM) to solve synonym problems [26].

In general, cross-language patent retrieval system can be developed by two approaches. In the first type of techniques, researchers used machine-translation based approaches to translate texts and queries into target languages for patent retrieval[13][9][23]. Utiyama and Isahara [33] utilized machine translation to construct a Japanese-English patent parallel corpus provided by the NTCIR-6 patent retrieval task. On the other hand, some language independent methods were developed to tackle the issues. In such solutions documents and queries are mapped into a semantic space to achieve cross-language retrieval. For instance, Dumais [4] proposed Cross-language Latent Semantic Indexing (CL-LSI) applying SVD process on English-French document. Li and Shawe-Taylor [21][22] utilized the kernel canonical correlation analysis (KCCA) method to retrieval patents in different languages.

To cope with the complexity of today's patent retrieval tasks, many hybrid methods have been developed to enhance the traditional approaches; for instance, Kang [14] utilized text mining techniques and a language model to the invalidity search task of patent retrieval. Uchida [32] proposed a generation-patent-map system based on a vector space model and clustering techniques to produce good cluster labels. Lamirel [18] used the self-organizing maps (SOM) [17] algorithm to cluster patent documents automatically. Tseng [30][31] used a multi-stage clustering method to support patent analysis, which is utilized to gradually identify the knowledge structures.

In addition, some researchers used ontology approaches to take advantages of common sense knowledge for patent information processing, including retrieval, interpretation, and

analysis of patent materials. For instance, Kitamura [16] developed an ontological framework to distinguish features of patent in each level for patent analysis, and Giereth [11] proposed a complex ontological framework for the representation of patent documentation.

In order to further to find the potential utility of patent documents' description sections, text mining techniques have been used to analyzing contents of patents, and to measure relatedness/similarity among patent documents. Based on the text mining techniques, Tseng [29] created a real world patent map for an important technology domain: "carbon nanotube", to reduce the research and development (R&D) time by "carbon nanotube" industry. Lee [20] utilized text mining technique to discover meaningful implications from the patent data, and used the results of patent analysis to finding business opportunities. Furthermore, text mining techniques also used to support patent analysis for the purposes of patent retrieval [8], summarization [37], and trend analysis [35][36].

However, results of patent analysis are normally not so comprehensible for users of non-specialists. Hence, the research of visualization methods for patent information and its analysis results has been attracted lots of attentions [5][7][24]. Camus and Brancaleon [2] developed a system (ArchIPat) to create, manage and analyze off-line patent collections. It enables a decision maker to analyze this knowledge and to build up a successful strategy. Hall [12] utilized visualization methods to show the number of patents and the number of patent citations, which significantly affect the market value of the company, and Kim [15] proposed a method for patent analysis, based on a semantic network of keywords extracted from patent documents. It is aimed to allow users easily understand the advances of emerging technologies and forecast its trend in the future.

On the other hand, some research work was concerned with patent classification tasks. Fall [10] applied a variety of machine learning algorithms for training expert systems (in German-language) to perform patent classification tasks. Teappey [28] developed a platform for patent document classification and search using a back-propagation neural network. Tikk [27] proposed a hierarchical online classifier algorithm to support categorizing patent applications in the IPC hierarchy.

V. CONCLUSION

In this research, we established a unified vector space model to allow patents in different languages map into a single document space and developed a method to evaluate the relatedness/similarity among multilingual patent documents by means of text clustering techniques. The process of synonym elimination has been considered to reduce the side effects of sparse matrix in LSI, and similar multilingual patent documents will be grouped by clustering process. Both recall and precision are considered in balance to choose the best experimental results. This clustering approach has been verified to help multilingual patent retrieval in our system framework.

Two issues need to be further addressed. First, there are many related technologies involved in the retrieval and analysis of patent claims; however, most of them are not belonging to the same IPC code, and it will affect clustering performance of the system, including recall and precision measures. Second, the collected training and test samples selected were focusing on the fields of semiconductor and optics. For future work, it is applicable to employ our developed platform to more professional domains.

REFERENCES

[1] R. Attar and A.S. Fraenkel, "Local feedback in full-text retrieval systems", Journal of the Association for Computing Machinery 24(3), pp. 397–417, 1977.

[2] C. Camus and R. Brancaleon, "Intellectual assets management: From patents to knowledge", World Patent Information, 25(2), pp.155–159, 2003.

[3] S. Deerwester, S.T. Dumais, G.W. Furnas, T. K. Landauer, and R. Harshman, "Indexing By Latent Semantic Analysis", Journal of the American Society for Information Science, Vol.41, pp. 391-407, 1990.

[4] S.T. Dumais, and T.L. Landauer, and M.L. Littman "Automatic cross-language information retrieval using latent semantic indexing", SIGIR'96 Workshop on Cross-Linguistic Information Retrieval, 1996.

[5] H.J.-M. Dou, "Benchmarking R&D and companies through patent analysis using free databases and special software: A tool to improve innovative thinking", World Patent Information, 26, 297–309, 2004.

[6] European Patent Office, "Insufficient use of innovation support mechanisms in Europe", Information for Journalists, 2003. from http://www.epo.org/about-us/press/releases/archive/2003/05112003.html

[7] M. Fattori, G. Pedrazzi, and R. Turra, (2003). "Text mining applied to patent mapping: A practical business case", World Patent Information, 25(4), pp.335–342, 2003.

[8] A. Fujii, M. Iwayama, N. Kando, "Introduction to the special issue on patent processing", Information Processing & Management, 43(5), pp.1149–1153, 2007.

[9] A. Fujii, M. Utiyama, M. Yamamoto, and T Utsuro, "Overview of the Patent Translation Task at the NTCIR-7Workshop", Proceedings of the 7th NTCIR Workshop, 2008.

[10] C.J. Fall, A. Törcsva'ri, P. Fie'vet, and G. Karetka, "Automated categorization of German-language patent documents", Expert Systems with Applications, 26(2), pp.269–277, 2004.

[11] M. Giereth, S. Koch, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, L. Serafini, L. Wanner und T. Ertl, "A Modular Framework for Ontology-based Representation of Patent Information", in: Legal Knowledge and Information Systems - JURIX 2007, Frontiers in Artificial Intelligence and Applications, vol. 165, pp. 49-58, IOS Press, 2007.

[12] B.H. Hall, A. Jaffe, and M. Trajtenberg, "Market value and patent citations", Rand Journal of Economics, 36(1), pp.16-38, 2005.

[13] S. Higuchi, M. Fukui, A. Fujii, and T. Ishikawa, "PRIME: A system for multi-lingual patent retrieval", In Proceedings of MT Summit VIII, pp. 163–167, 2001.

[14] I.S. Kang, S.H. Na, J. Kim, and J.H. Lee, "Cluster-based patent retrieval", Information Processing & Management, Vol. 43, Issue 5, pp. 1173-1182, 2007.

[15] Y. Kim, J. Suh, and S. Park. Visualization of patent analysis for emerging technology. Expert Systems with Applications, Vol. 34 (3), pp 1804-1812, 2007.

[16] Y. Kitamura, M. Kashiwase, M. Fuse, and, R. Mizoguchi, "Deployment of an Ontological Framework of Functional Design Knowledge", *Advanced Engineering Informatics*, 18(2), pp115-127, 2004.

[17] T. Kohonen, "Self-organizing maps", Berlin: Springer-Verlag, 1995.

[18] J.C. Lamirel, S.A. Shehabi, M. Hoffmann, and C. François,

"Intelligent Patent Analysis through the Use of a Neural Network: Experiment of Multi-Viewpoint Analysis with the MultiSOM Model", Proceedings of ACL Workshop on Patent Corpus Processing, pp. 7-23, 2003.

[19] L.S. Larkey, "A patent search and classification system", In Proceedings of the fourth ACM conference on digital libraries, pp. 179–187, 1999.

[20] E. Lee, B. Yoon, C. Lee, and J. Park, "Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping", Technological Forecasting and Social Change, 2009.

[21] Y. Li and J. Shawe-Taylor, "Using KCCA for Japanese-English Cross-Language Information Retrieval and Classification", In Journal of Intelligent Information Systems, 2005.

[22] Y. Li and J. Shawe-Taylor, "Advanced learning algorithms for cross-language patent retrieval and classification", Information Processing and Management, 43(5), p.p1183–1199, 2007.

[23] M. Makita, S. Higuchi, A. Fujii, and T. Ishikawa, "A system for Japanese/English/Korean multilingual patent retrieval", In Proceedings of Machine Translation Summit IX, 2003.

[24] S.Morris, C. DeYong, Z. Wu, S. Salman, and D. Yemenu, "DIVA: A visualization system for exploring documents databases for technology Forecasting", Computers and Industrial Engineering, 43(4), 841–862, 2002.

[25] M. Osborn, T. Strzalkowski, and M. Marinescu, "Evaluating document retrieval in patent database: a preliminary report", In Proceedings of the conference on information and knowledge management, pp. 216–221, 1997.

[26] J.F. Ryley, J. Saffer, and A. Gibbs, "Advanced document retrieval techniques for patent research", World Patent Information, pp. 238-243, 2008.

[27] D. Tikk, G. Biró, and A. Törcsvári. "A Hierarchical Online Classifier for Patent Categorization", In H. A. do Prado and E. Ferneda, editors, Emerging Technologies of Text Mining: Techniques and Applications. Idea Group Inc., 2007.

[28] A.J.C. Trappey, F.C. Hsu, C.V. Trappey, and C.I. Lin, "Development of a patent document classification and search platform using a back-propagation network", Expert Systems with Applications, 31(4), pp.755–765, 2006.

[29] Y.H. Tseng, Y.M. Wang, D.W. Juang, and C.J. Lin, "Text mining for patent map analysis", In IACIS Pacific 2005 conference proceedings, pp. 1109–1116, 2005.

[30] Y.H. Tseng, C.J. Lin, and Y.I. Lin, "Text Mining techniques for patent analysis", Information Processing and Management, Vol. 43, No.5, pp. 1216-1247, 2007.

[31] Y.H. Tseng, Y.M. Wang, Y.I. Lin, C.J. Lin, and D.W. Juang, "Patent Surrogate Extraction and Evaluation in the Context of Patent Mapping", In Journal of Information Science, 2007.

[32] H. Uchida, A. Mano, and T. Yukawa, "Patent map generation using concept-based vector space model", In Proceedings of the fourth NTCIR workshop, 2004.

[33] M. Utiyama and H. Isahara, "A Japanese-English patent parallel corpus", In Proceedings of MT Summit XI, pp. 475–482, 2007.

[34] T. Yeap, G. Loo, and Pang, S., "Computational patent mapping: intelligent agents for nanotechnology", Proceedings of the International Conference on MEMS, NANO and Smart Systems, pp. 274–278, 2003.

[35] T. Yeap, G. Loo, and S. Pang, "Computational patent mapping: intelligent agents for nanotechnology", Proceedings of the International Conference on MEMS, NANO and Smart Systems, pp. 274–278, 2003.

[36] B. Yoon, Y. Park, A text-mining-based patent network: analytic tool for high-technology trend, J. High Technol. Managem. Res, 15(1), pp.37–50, 2004.

[37] B. Yoon, R. Phaal, and P. David, "Structuring technological information for technology roadmapping: Data mining approach", Proceedings of the 7th WSEAS International Conference on Artificial intelligence, knowledge engineering and data bases, 2008.