A Text-Mining-based Patent Analysis in Product Innovative Process

Liang Yanhong[1], Tan Runhua[2]
Institute of Design for Innovation, Hebei University of Technology,
Tianjin, 300130, P.R. China
Email: 1 waterlily00@126.com
2 rhtan@hebut.edu.cn

Abstract. Patent documents contain important technical knowledge and research results. They have high quality information to inspire designers in product development. However, they are lengthy and have much noisy results such that it takes a lot of human efforts for analysis. And due to the fact that hidden and unanticipated information plays a dominant role for TRIZ user, it is difficult to discern manually, thus, patent analysis has long been considered useful in product innovative process. Automatic tools for assisting innovators and patent engineers in obtaining useful information from patent documents are in great demand. In TRIZ theory, a product design problem can be considered as one or several Contradictions and Inventive Principles. Text mining could be used to analyze these textual documents and extract useful information from large amount documents quickly and automatically. In this paper, a computer-aided approach for extracting useful information from patent documents according to TRIZ Inventive Principles is proposed.

Keywords. patent analysis, TRIZ, Inventive Principles, text mining

1 Introduction

A major competitive advantage for any company is the ability of product innovation. With the aid of massive information across the globe, highly complex products are needed to develop to meet the customer needs at a very low cost but in ever-shorter time. Product development is a process, which builds on the basis of knowledge and experience to solve problems. The technical goal of product development is to reduce difference between the initial state and the idealist state of the product, and to look for the best scheme to improve product quality and reliability. Accompany with increasing complexity of products in modern society, the traditional trial is in low efficiency and brainstorming becomes unreliable to product innovation.

Please use the following format when citing this chapter:

Yanghong, L., Runhua, T., 2007, in IFIP International Federation for Information Processing, Volume 250, Trends in Computer Aided Innovation, ed. León-Rovira, N., (Boston: Springer), pp. 89-96.

Altshuller, the father of TRIZ, recognized that the place where to look for the basics of innovation and new ideas was not in the brains of inventors, but in the published inventions [1]. In reviewing thousands of patents, Altshuller and his colleague categorized the inventive principles in several retrievable forms, including a contradiction table, 39 Engineering Parameters, 40 Inventive Principles, and 76 Standard Solutions [2]. He provided a systematic process to define and solve any given problem, especially in the field of product development. According to TRIZ, a significant operation of product innovation is to solve design contradictions. When contradictions and Inventive Principles are defined, product is developed by referring to the analogous inventions not only in related fields but also dissimilar problems in other fields that have previously solved the same contradiction.

In recent years, patent analysis [3,4,5] is more highlighted in high-technology management as the process of innovation becomes more complex, the cycle of innovation becomes shorter and the market demand becomes more volatile. Patent documents contain important research results that are valuable for product innovation. However, they are lengthy and have much noisy results such that it takes a lot of human efforts for analysis manually. To obtain useful information, the method by scanning or reading the indexed patent documents from long lists of noisy results, is a rather trivial and time-consuming task that requires a careful manual selection. And the defects of extract information from patent documents, which indexed by standard keyword-based search methods, will ignore relevant solvable schemes and enlightenment in other fields. In addition to the huge requirement of manpower and time, the rapid increase of the number and application of patent documents, thus, there is a need to find a way to get ride of tradeoffs and obtain useful and precise patent documents quickly. One method to solve this difficult problem is data mining. Data mining is an automated scheme to extract useful information from large databases. As to patent documents are nearly unstructured texts, text mining, like data mining or knowledge discovery, which specialized for full-text patent analysis, is applied to derive information.

Thus, automatic tools of text mining in patent analysis for assisting innovators or patent engineers are in great demand.

Further elaboration on text mining and patent classification would be provided in section 2. Following that, elaboration on the background of text mining for patent analysis would be provided in the rest of paper. Following that, a text-mining approach that helps extract and analyze useful information automatically from online network-patent documents is discussed. Thus, automatic tools and softwares for assisting TRIZ users and patent engineers to select useful patent documents from large amount of patents are given. Further, some of the difficulties in extracting information not only concise but also not loss are also presented. Finally, future research work is discussed. Our focus, however, would be largely directed towards the methodology to extract textual components from patent documents.

2 Theoretical background

2.1 Patent classification and characteristics of patent document

Patent classification schemes are used to organize and index the technical content of patent specifications so that specifications on a specific topic or in a given area of technology can be identified easily and accurately. Before their publication, patent documents are given one or more classification codes based on their textual contents for topic-based analysis and retrieval. Many patent classification schemes, such as IPC (International Patent Classification), US Classification and British Classification, have been developed. However, the classification schemes used by these researchers are based on the application fields involved in the inventions, For example, the IPC divides patent technology into 8 key areas:

A: Human Necessities
B: Performing Operations, Transporting
C: Chemistry, Metallurgy
D: Textiles, Paper
E: Fixed Constructions
F: Mechanical Engineering, Lighting, Heating, Weapons
G: Physics
H: Electricity

Patent documents are divided into different areas according to technology fields is helpful to search the prior art for traditional inventors. However, it is inadequate for TRIZ users since TRIZ users are interested in previous patents that have solved the same Contradiction and used the same Inventive Principles, which may come from different fields [6,7]. Patent documents, which are classified according to Inventive Principles combined with Contradiction, will provide a broader view for TRIZ users and TRIZ software developers, by helping them find possible inspiration from a field that may be totally different from theirs.

A patent document contains dozens of items for analysis. Some are structured, that is to say they are uniform in semantics and in format such as patent number, filing date, or assignees; some are unstructured, that is to say they are free texts of various lengths and contents, such as title, claims, abstract, or descriptions of the invention. The description of the invention can be further segmented into field of the invention, background, summary, and detailed description, although some patents may not have all these segments. Patent analyses based on structure information have been the major approaches in practice and in the literature for years [8,9,10]. These structured data can be analyzed by data mining techniques or well-established database management tools such as OLAP (On-Line Analytical Processing) modules. But the most rest of patent document is made of unstructured text, based on this, there has been an interest in applying text mining techniques to assist the task of patent analysis. Based on Text mining people do not have to understand text in order to extract useful information from it.

2.2 Text mining

Data mining, also known as knowledge discovery in a database, is a recent development for accessing and extracting information in a database [11,12]. In short,

data mining applies machine-learning and statistical analysis techniques for the automatic discovery of patterns in a database. Most efforts in data mining, however, have been made to extract information from a structured database and the utility of data mining is yet limited in handling huge amounts of unstructured textual documents.

As a remedy, text mining [13,14] is a rather new technique that has been proposed to perform knowledge discovery from collections of unstructured text. Like data mining or knowledge discovery, text mining is often regarded as a process to find implicit, previously unknown, and potentially useful regularities in large textual datasets. In briefly speaking, text mining puts a set of labels on each document, and discovery operations are performed on the labels. The usual practice is to put labels to words in the document. Then, the document in text format can be featured by keywords and clue words that are extracted through text mining algorithm. Since it is suitable for drawing valuable information from large volumes of unstructured text, text mining has been widely adopted to explore the complex relationship among patent documents. With the application of text mining an effective means can be provided for content searches in the textual fields of patent documents.

3 A text-mining-based methodology to patent analysis

Based on the theoretical background introduced above, a general methodology is suggested to analyze patent documents. The overall process of conducting text-mining-based patent analysis goes through several steps. First of all, text collection and text preprocessing are the preliminary step. The interested patent area is selected and related patent documents are collected in electronic text format. Second, raw patent documents are transformed into structured data. Since the original documents are expressed in natural language format, they must be transformed into structured data in order to be analyzed and utilized. Text mining that extracts keywords and clue words from patent document is used to this end. Fig.1 depicts the overall process of text-mining-based patent analysis [15]. In relation to patent analysis, text mining is used as a data processing and information-extracting tool. Since the original patent documents are expressed in natural language format, it is necessary to transform raw data into structured data. Then, the process of keyword and clue word extraction is applied to measure similarity between patents.

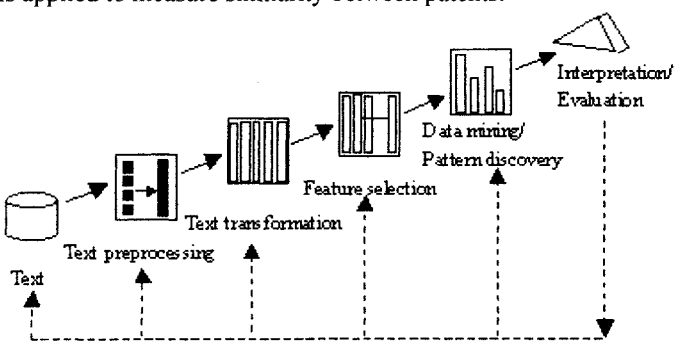

Fig.1. Text mining process of patent analysis

3.1 Text preprocessing

The patent documents in our experiment from USPTO(United States Patent and Trademark Office: www.uspto.gov) are in HTML format and contains title, abstract, claims, and description. When manually analyze the documents, we found that usually the abstracts and descriptions provided enough semantic information to determine TRIZ Inventive Principles that the patents used. Therefore the abstracts and descriptions are extracted from the full text and parsed into independent sentence. As pointed out above, raw documents need to be processed because they are unstructured in format. The description, the main body of the detailed content, often have sub-sections with titles in uppercase, such as FIELD OF THE INVENTION, BACKGROUND, SUMMARY OF THE INVENTION, and DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT. Although some patents may have more or fewer such sub-sections or have slightly different title names, most patents do follow this style. Thus a regular expression matcher is devised to extract each of these segments. It is implemented using regular expressions in Perl. The method takes advantage of the rule that each sub-section's title is in a single line paragraph separated by two HTML tags: "

". After splitting the paragraphs based on these tags, a set of Perl expressions: (/Abstract/i, /Claims/i, /FIELD/i, /BACKGROUND/i, /SUMMARY/I, /DESCRIPTION|EMBODIMENT/i) are used to match the patent segments.

3.2 Text transformation and feature selection

The document is segmented at word level as the smallest unit. The document is first split into a series of words. Each document is made of bags of words. Adjectives, adverbs, nouns and multi-word are extracted from the document. Word frequency (term frequency) and inverse document frequency are two parameters used in filtering terms. Low TF and DF terms are often removed from the indexing of patent documents. After removing stop words [16] in each document, word stemming [17] is performed. To better match concepts among terms, words are stemmed based on Porter's algorithm [18]. It contains keywords, title words, and clue words. Here the keywords are those maximally repeated words that can be calculated by a fast key term extraction algorithm [19]. The algorithm works with the help of a stop word list alone. By repeatedly merging back nearby words based on three simple merging, dropping, and accepting rules, Maximally repeated strings in the text are thus extracted as keyword candidates. The algorithm is shown in fig.2. The title words are those non-stop words that occur in the title of a patent document. As to the clue words, they are a list of about 25 special words that reveal the intent, functions, purposes, or improvements of the patent. These words are prepared by several patent analysts based on their experiences and are listed as table 1.

```
For each patent document
{
    1.1 Convert the text into a LIST of words
    1.2 Do
    {
      2.1 Set MergeList to empty
      2.2 Put a separator to the end of LIST as a sentinel and set the occurring frequency of the separator to 0
      2.3 For I from 1 to NumOf(LIST) - 1 step 1
          {
            3.1 If LIST[ I ] is the separator, then
                  Go to Label 2.3.
             3.2 If Freq(LIST[ I ]) > threshold and Freq(LIST[I+1]) > threshold, then
                  Merge LIST[ I ] and LIST[I +1] into Z
                  Put Z to the end of MergeList
                Else
                  If Freq(LIST[ I ]) > threshold and LIST[ I ] did not merge with LIST[ I - 1], then
                    Save LIST[ I ] in FinalList.
                  If the last element of MergeList is not the separator, then
                    Put the separator to the end of MergeList.
          }
      2.4 Set LIST to MergeList.
    }while NumOf(LIST) < 2
    1.3 Filter terms in FinalList based on some criteria
}
```

Fig.2. The keyword extraction algorithm

Table 1. the clue words for the BACKGROUND segment

Advantage	Difficult	Improved	Overhead	Shorten
Avoid	Effectiveness	Increase	Performance	Simplify
Cost	Efficiency	Issue	Problem	Suffer
Costly	Goal	Limit	Reduced	Superior
Decrease	Important	Needed	Resolve	Weakness

3.3 Pattern discovery

Whenever the designer faces a problem in product development, he would search for inventive experience and information to enlighten his thinking. However, this search is not an easy one to TRIZ user. A keyword search is helpful to the innovator but sometimes the useful patents maybe be neglected since the same Inventive Principles to solve the problem come from different fields. Therefore there is a need to accurately return records that might exhibit similar problems and causes to innovators. Under these circumstances, clustering might be a possible tool that could be employed. Records could be initially placed into groups using clustering algorithm [20] and multi-Naïve Bayes algorithm [21,22,23]. Then, as TRIZ user

faces a new problem, that is to say, when an Inventive Principle is given, records that exhibit similar problems to the groups could be returned.

5 Conclusion and future work

This paper describes characteristics of patent document and text mining technique. A methodology based text mining that could be used to analyze patent document for TRIZ user is presented. It points out clustering algorithm and multi-Naïve Bayes algorithm to extract useful information from patent document, which is helpful in product innovative process for TRIZ users. However, more work is required and some difficulties exist. Firstly, to put the methodology into practice, an automated text mining system will be developed using Perl and WEKA software [24]. Secondly, in addition to TRIZ Incentive Principles, Contradictions would also be taken into consideration. The difficulty is that the textual dataset of patent is not a numerical dataset. Users in different circumstance, or with different needs, knowledge of linguistic habits will describe the same information using different terms. What is more, sometimes the same word could have more than one distinct meaning. These add to difficulties of understanding the texts.

6 Acknowledgement

The research is supported in part by the Chinese national 863 planning project under Grant Number 2006AA04Z109. No part of this paper represents the views and opinions of any of the sponsors mentioned above.

7 References

1. R.H. Tan, *Theory of Inventive Problem Solving: TRIZ* (Science Press, China, 2004).
2. http://www.oxfordcreativity.co.uk/ (April 10, 2007)
3. V.W Soo, S.Y Lin, S.Y Yang, S.N Lin, S.L Cheng, A cooperative multi-agent platform for invention based on patent document analysis and ontology, *Expert Systems with Applications 31* (2006), pp766-775
4. B. Yoon, Y. Park, A text-mining-based patent network: Analytical tool for high-technology trend, *Journal of High Technology Management Research 15*(2004), pp. 37-50.
5. G. Fischer, N. Lalyre, Analysis and visualization with host-based software-The features of STN AnaVist, *World Patent Information 28*(2006), pp. 312-318.
6. H.T. Loh, C. He, L.X. Shen, Automatic classification of patent document for TRIZ users, *World Patent Information 28*(2006), pp. 6-13.
7. C. He, H.T. Loh, Grouping of TRIZ Inventive Principles to facilitate automatic patent classification, *Expert System with Applications* (2006), pp1-8
8. D. Archibugi, M Pianta, Measuring technological change through patents and innovation survey, *Technovation*, 16(9) (1996) , pp.451-468.
9. B.D.Carrax, C. Hout, A new methodology for systematic exploitation of technology databases, *Information Processing & Management*, 30(3) (1994), pp407 - 418.

10. K.K Lai, S.J. Wu, Using the patent co-citation approach to establish a new patent classification system, *Information Processing & Management*, 41(2) (2005), pp313 - 330.
11. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery In Databases, *American Association for Artificial Intelligence* (1996), pp37-56.
12. G. Piatetsky-Shapiro, U. Fayyad, P. Smyth, Advances in knowledge discovery and data mining, *AAAI/MIT Press* (1996), pp1–33.
13. I.H. Witten. Text mining, *Practical handbook of Internet computing* (CRC Press, Boca Raton, 2004).
14. Witten I.H. Bray. Z, Mahoui. M, and Teahan .W, Text mining: A new frontier for lossless compression. *Proceedings of the Data Compression Conference,* Snowbird, UT.Los Alamitos, CA:IEEE Computer Society Press (1999), pp.198-207.
15. A. Wasilewska, cse634-Data Mining: Text Mining (June 5, 2007); http://www.cs.sunysb.edu/~cse634/presentations/
16. Onix Text Retrieval Toolkit (June 5, 2007); http://www.lextek.com/manuals/onix/stopwords1.html.
17. Porter Stemming Algorithm (March 5,2007); http://www.tartarus.org/~martin/PorterStemmer/.
18. M.F. Porter, An algorithm for suffix stripping, *Program* (1980), 14(3), pp130-137.
19. Y.H. Tseng, C.J. Lin, Y.I Lin, Text mining techniques for patent analysis, *Information Processing and Management* (2007)
20. A. Gatt, Structuring Knowledge for Reference Generation: A Clustering Algorithm (March 2, 2007); http://www.csd.abdn.ac.uk/~agatt/home/pubs/
21. Y. Yang, X. Liu, A re-examination of text categorization methods, *SIGIR99* (1999).
22. I.H. Witten, E. Frank, Data Minging: *Pritical Machine Learning Tools and Techniques, Second Edition* (Elsevier Inc, Singapore, 2005).
23. A. McCallum, K. Nigam, A comparison of event models for Naïve Bayes text classification, *Proceedings of the AAAI-98 Workshop on Learning for Text Categoration, Madison, WI.Menlo Park,* CA:AAAI Press(1998), pp41-48.
24. WEKA (September 2, 2006); http://www.cs.waikato.ac.nz/ml/weak/.