Engineering Applications of Artificial Intelligence 24 (2011) 1510–1520

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis

W.M. Wang*, C.F. Cheung

Knowledge Management Research Centre, Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 20 June 2010
Received in revised form
6 April 2011
Accepted 14 May 2011

Keywords:
Knowledge management
Knowledge-based system
Semantic analysis
Technology management
Concept extraction
Patent analysis

ABSTRACT

Patent databases provide valuable information for technology management. However, the rapid growth of patent documents, the lengthy text and the rich of content in technical terminology, and the complicated relationships among the patents, make it taking a lot of human effort for conducting analyses. As a result, an automated system for assisting the inventors in patent analysis as well as providing support in technological innovation is in great demand. In this paper, a Semantic-based Intellectual Property Management System (SIPMS) has been developed for supporting the management of intellectual properties (IP). It incorporates semantic analysis and text mining techniques for processing and analyzing the patent documents. The method differentiates itself from the traditional technological management tools in its knowledge base. Instead of eliciting knowledge from domain experts, the proposed method adopts global patent databases as sources of knowledge. The system enables users to search for existing patent documents or relevant IP documents which are related to a potential new invention and to support invention by providing the relationships and patterns among a group of IP documents. The method has been evaluated by benchmarking with the performance against traditional text mining technique and has successfully been implemented at a selected reference site.

1. Introduction

In knowledge-based economy, the management of intellectual properties (IP) has become more and more important for an industry. It also plays an important role in technology management. For instance, a patent is a contract between an inventor and the government, whereby in return for full public disclosure of an invention, the government grants the inventor the right to exclude others for a limited time from making, using and selling the invention (Hufker and Alpert, 1994). Many companies hold their patents not only as an invisible asset, but also as a strategy for its development and competition against its competitors in the market (Jung, 2003). However, inventing a new patent is not an easy task. Some "novel" ideas might impinge on some claimed rights protected by other's patents. Therefore, before inventing a new patent, it should ensure that it does not infringe other's patents by consulting with some global patent databases. In fact, the World Intellectual Property Organization revealed that from 90% to 95% of world's inventions are found in patented documents (Brockhoff et al., 1999). The European Patent Office also disclosed that more than 80% of man's technical knowledge is described in patent literature (European Patent Office, 2010). By carefully analyzing the patent documents, it cannot only increase the efficiency and effectiveness of making new invention substantially but also reduce the risk of infringing the patent rights of others (Soo et al., 2005).

Patent documents analysis has often been employed to generate economic indicators that gage the linkage between technology development and economic growth (Campbell, 1983; Grandstrand, 1999; Grilliches, 1990; Holl et al., 2000), estimate technological knowledge flows and their impact on productivity (Evenson and Puttnam, 1988; Scherer, 1982), compare innovative performance in international context (Paci et al., 1997), evaluate the competitiveness of firms (Narin and Noma, 1987), develop technology plans (Mogee, 1991), prioritize R&D investment (Hirschey and Richardson, 2001), or monitor technological change in firms (Archibugi and Pianta, 1996; Basberg, 1987). However, patent analysis requires considerable effort and expertise. It requires the analysts to have a certain degree of expertise in information retrieval, domain-specific technologies, and business intelligence. In addition, patent documents are often lengthy and rich in technical and legal terminology. It consumes a lot of time to read and analyze them even for experts. As a result, it is necessary to have an automated method for supporting the analysts in processing and analysis of massive patent documents as well as supporting technology innovation.

* Corresponding author.
E-mail address: mfming@inet.polyu.edu.hk (W.M. Wang).

0952-1976/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.05.009

The paper is outlined as follows. Section 2 reviews the related work of the study which includes patent analysis, text mining, and keyword extraction. Section 3 describes the architecture and the components of the proposed system. Experiments and results for evaluating the system are provided in Section 4. The usefulness and effectiveness of the system in a real-life environment have been studied in Section 5 through a practical implementation of a prototype at a selected reference site. Section 6 is an overall conclusion of the paper, and some suggested areas for further work are also explored in this section.

2. Related work

Most studies in the patent analysis are on document search and classification (Fall et al., 2003; Larkey, 1999; Lai and Wu, 2005; Tsakalidis et al., 2002). A patent document contains a lot of items for analyses. Some of them are structured data. They are uniform in semantics and in format across patents such as patent number, filing date, or assignees. Some are unstructured data. They are free texts of various lengths and contents, such as claims, abstracts, or descriptions of the invention. Patent analyses based on structured information have been most practised approaches and have been found in the literature for years (Archibugi and Pianta, 1996; Be'de'carrax and Huot, 1994; Ernst, 1997; Lai and Wu, 2005).

These structured data can be analyzed by bibliometric methods, data mining techniques, or well-established database management tools such as On-Line Analytical Processing (OLAP). One type of structured data analyses is determined by citations (Karki, 1997). For example, if a patent is cited by a large number of other patents, this cited patent is possibly a foundation of those citing patents and is thus important. In another example, since two patents related to the same invention tend to cite the same patent and are cited by the same patent, interpatent similarity can be determined by cocitation analysis. Using patent citation analysis, Narin et al. (1984) evaluated corporate technological performance while Chakrabarti et al. (1993) analyzed the diffusion of technological information among different organizations. Lai and Wu (2005) used citation-based interpatent similarity to perform a patent classification. Recently, there has been an interest in applying text mining techniques to assist the task of patent analysis and patent mapping (Lent et al., 1997; Fattori et al., 2003; Yoon and Park, 2004).

Text mining can be viewed as data mining extended to text data to find implicit, previously unknown, and potentially useful patterns from a large text repository (Fayyad et al., 1996). It is an interdisciplinary area involving machine learning and data mining, statistics, information retrieval and natural language processing (NLP) (Grobelnik et al., 2002). Text mining can work with unstructured datasets such as full-text documents, HTML files, emails, etc. NLP techniques are commonly used as the first step in text mining for converting the unstructured text into a structured format (Milic-Frayling, 2005). The document can be featured by keywords that are extracted through text mining algorithms, and then other data mining techniques can be applied to retrieve interesting patterns. In relation to patent analysis, text mining is used as a data-processing and information-extraction tool. Since the original patent documents are expressed in text (natural language) format, it is necessary to transform raw data into structured data. Then, the process of keyword extraction is employed to identify keywords and to measure similarity between patents.

In the present study, a Semantic-based Intellectual Property Management System (SIPMS) is presented which attempts to address the shortcoming of the prior research in several respects. It applies text mining to analyze the unstructured part of the patent documents. It extracts the key concepts of the patent documents and discovers the relations among the concepts based on the syntactic structure of the documents. This approach is able to extract concepts and relationships from the document itself rather than retrieved from predefined ontologies. Hence, it enables the approach to be applied in any domains without the need for the capture of prior knowledge. It also ensures that the intentions of the author of the document can be preserved using the words generated from the document itself. Furthermore, the proposed algorithm produces concepts and relationships based on multi-word text structures instead of individual words which makes the concept and relationship labels to be more complete. Based on the proposed approach, the results of classification of the patents can achieve a higher accuracy.

Prior research into keyword extraction is summarized in three categories, which are dictionary (e.g. Braam and Moed, 1991; Callon et al., 1991; Zitt and Bassecoulard, 1994), statistical (e.g. Cutting et al., 1992; Eisen et al., 1998; Karypis et al., 1994), and linguistic approaches (e.g. Rajaraman and Tan, 2002; SanJuan and Ibekwe-SanJuan, 2006).

The dictionary approach utilizes a dictionary, which contains the forms, meanings and relationships between words and phrases. By matching the dictionary with the words of sentences in each article, the concepts in the article are extracted. Relationships between the concepts are then coordinated based on the dictionary. The major advantages of this approach are its efficiency of execution and ease of implementation. However, the prior keyword list is external to the documents. It reflects a general understanding of the domain instead of the intention of the author of the document. Since words in the target articles cannot consist of new words, it requires further operations for handling these new words. In the statistical approach, words are selected based on term weighting indices such as Inverse Document Frequency (IDF) or Mutual Information (MI). It also eliminates the low frequency words so as to reduce the number of words being extracted. However, this also results in the drastic elimination of more than half of the initial data from the analysis. Moreover, Price and Thelwall (2005) have demonstrated the usefulness of low frequency words for scientific web intelligence (SWI). Removal of low frequency words results in documents becoming more general and similar. The linguistic approach makes use of semantic knowledge bases, heuristics, or rules to extract concepts. However, from a conventional linguistic approach, it is hard to make linguistic generalizations that can be applied reliably due to the occurrence of ambiguous words and ambiguous sentences structures.

In addition, patent analysis is a complex task which consists of different objectives or goals. A multi-agent system (MAS) architecture has been developed which consists of a collection of autonomous agents which have defined their own goals and actions and can interact and collaborate among each other through XML communication. These agents act collectively and collaborate to achieve their own individual goals as well as the common goal.

3. The architecture of Semantic-based Intellectual Property Management System (SIPMS)

The architecture of SIPMS is shown in Fig. 1. It is a knowledge-based system which consists of three major processes which are pre-processing, patent analysis, and invention support. The combination of the algorithms is presented here for the first time for supporting technological innovation.

3.1. Pre-processing

There are different kinds of intelligent agents in the processes. The pre-processing process consists of extraction agent, segmentation

agent, and indexing agent. This process aims at selecting the relevant patents, divide each patent into different sections and indexing the documents for further analysis. As shown in Fig. 2, based on the user inputted filtering criteria, such as International Patent Classification (IPC), the extraction agent keeps checking with the relevant patent documents from the external and internal patent databases. If there is a new patent, the agent extracts the patent based on a predefined patent schema. And it stores the patent document into the internal patent database. After that, the segmentation agent divides the selected patents into a semi-structured format which include filing date, application date, assignees, IPC codes, title, abstract, claims, and description of the invention. A regular expression matcher is devised to extract each of these segments.

The indexing agent converts the semi-structured format into concepts. It is accomplished by a concept extraction algorithm purposely developed by the authors. Generally, the subject of a sentence represents a concept. The object of a sentence, which is in the verbal phrase, represents a second concept. The algorithm extracts all the noun phrases inside the text and consolidates them into a list of key concepts. The unstructured text is first divided into tokens by regular expressions such as the new line character, full stop, question mark, etc. The tokens are then tagged with their parts-of-speech (POS) using a POS tagger developed by Schmid (1994). The tags set is shown in the Appendix A. Each token is merged with its nearby token as noun phrases based on their POS. For example, if the nearby token's POS are "nn", "nns", "nnp", "nnps", "np", "prp", "prp$", "pp$", or preposition "of", then the token are merged with the noun phrase.

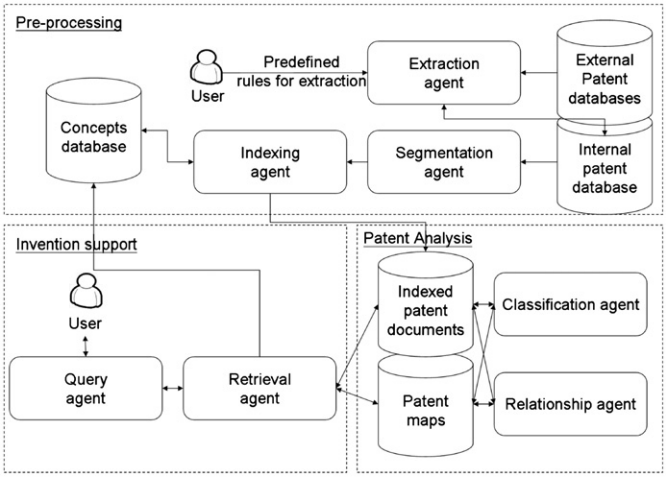

Fig. 1. The architecture of SIPMS for supporting invention.

The extracted concepts are then weighted by a hybrid weighting approach which is developed based on heuristic rules, semantic, and deep syntactic analysis. The method identifies the relationships among the extracted concepts so as to determine weights to them. It is assumed that a concept has a higher weight if the concept has more relationships with other concepts. In the present study, the heuristic rules analysis has been used to detect the simple syntactic patterns among tags in a timely and effective manner. Two basic heuristic rules are used as shown below:

(a) Rule 1: when one term is the same as the other term and additionally modified by certain words or adjectives, the longer term is always categorized as a part of the shorter terms. Thus, a rule to detect vertical relations by a simple is—a rule, which is adapted from the research of Velardi et al. (2001). Basically, given two terms c_1 and c_2, if c_2 matches c_1 and c_1 is additionally modified by certain words, then the relation relationship(c_1, c_2) is derived. For example, c_1='credit card' and c_2='card', the relation of the relationship('credit card', 'card') is derived.

(b) Rule 2: a rule to detect abbreviations. Given two terms c_1 and c_2, if c_2's alphabet letters matches the first letters of words of c_1, then c_2 is the abbreviation of c_1 and c_1 and c_2 are considered to have a neighbor relationship. For example, c_1='natural language processing' and c_2='NLP', the relation of the relationship('natural language processing', 'NLP') is derived. This rule serves to group the abbreviation under its original (expanded) phrase so as to minimize the complexity of the extracted concepts.

In semantic analysis, WordNet (Miller, 1995) provides common knowledge information to match words based on linguistic relations between them (e.g. synonyms, hyponyms). For instance, Alves et al. (2002) used WordNet to extract an initial hierarchy of nouns from a document to build an initial list of concepts, followed by several user feedback iterations to deduce relationships between pairs of concepts and hypothesize about their relations. Paik et al. (2001) defined a representation called concept–relation–concept (CRC) triples. Their system analyzes raw text to construct a database of CRC triples based on semantic relations. In the present study, the system adapts the dataset of WordNet 2.1. There are 19 pointer symbols for nouns for representing the relationships between 2 nouns (e.g. hypernym, meronym, etc.). They are all classified having relations in this research. For example, canines is a hypernym of dog. Then we consider canines and dog have a relationship.

The third part of the method is based on a concept–relationship acquisition and inference algorithm which is adapted from an automatic concept mapping algorithm (Wang et al., 2008). The algorithm converts raw text documents into a

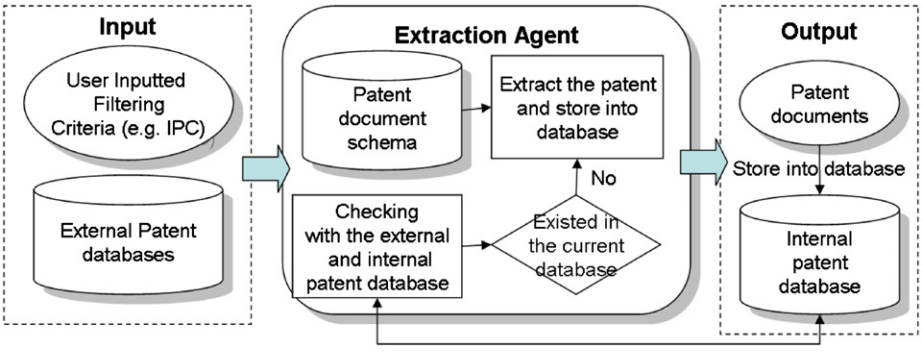

Fig. 2. A schematic diagram of the extraction agent.

"concept–relationship–concept" format. In the present study, concepts have a semantic relationship if they constructed a "concept–relationship–concept" format after the concept mapping algorithm analysis. For example, c_1='Content Management' and c_2='Content Protection' and the relationship between c_1 and c_2 is 'consists of' then, relationship(c_1, c_2) is derived.

After that, the weighting of each concept is determined by Eq. (1)

$$w_i = \sum_{i,j=1, i\neq j}^{n} \frac{relationship(c_i, c_j)}{n} \quad (1)$$

where w_i is the weight of concept i, $relationship(c_i, c_j)$ is the relationship between concept i and concept j, $relationship(c_i, c_j)=1$ if concept i and concept j consist relationship, $relationship(c_i, c_j)=0$ if concept i and concept j do not consist relationship, n is the total number of extracted concepts. The unstructured text is then indexed with its key concepts and their corresponding weightings.

3.2. Patent analysis

The analysis process consists of a classification agent and a relationship agent for generating patent maps. It aims at providing interesting patterns among a number of patent documents. Classification is a powerful technique to detect topics and their relations in a collection. Various classification methods have been proposed, such as Hierarchical Agglomerative Clustering (HAC) (Jain et al., 1999), Multi-dimensional Scaling (MDS) (Kruskal, 1977), and Self-organization Map (SOM) (Kohonen, 1997). Based on the indexed attributes of the patent documents, each document is processed into a vector form. In the present study, the patent classification agent makes use of a Naive Bayesian algorithm (Mozina et al., 2004) to form categories. The relationship agent associates the related concepts among the indexed patent documents. The algorithm is built based on a self-associated concept mapping (SACM) algorithm, which is developed by the authors (Wang et al., 2008). The graphical representation provides insights for describing the relationships among different knowledge concepts. A SACM is represented by a simple graph with nodes and edges. The nodes represent concepts relevant to a given domain and the association relationships between them are depicted by directed edges. An example of SACM is shown in Fig. 3. The importance of the concepts and the associations between different concepts are indicated by the depth of color i.e. darker color indicates higher importance.

SACM can be automatically constructed and dynamically updated from a knowledge repository with the indexed records. As shown in Fig. 4, the SACM consists of 3 major steps:

(i) *Step 1*: a temporary SACM is constructed based on the indexed text.
  – Distinct concepts are extracted from the text for the construction of a set of concept nodes **C** and the weight of concept W_i for each $C_i \in \mathbf{C}$ is assigned based on Eq. (1).

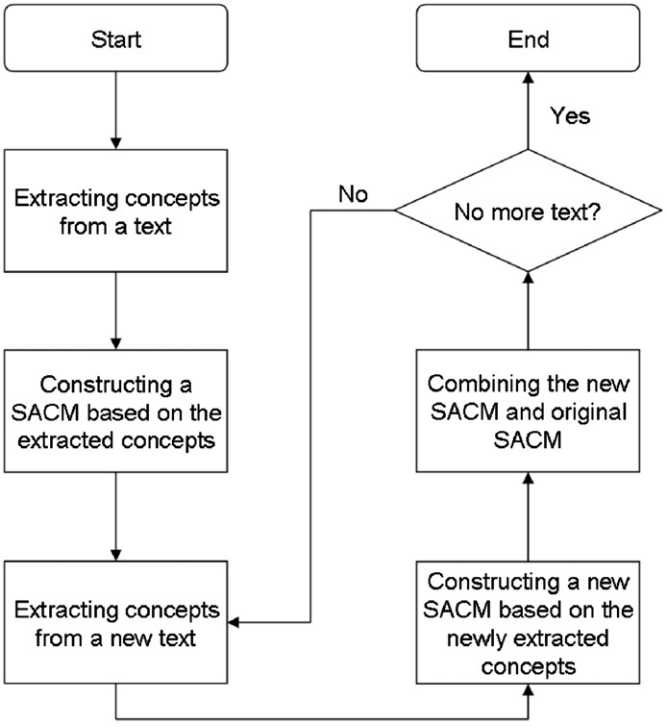

Fig. 4. A schematic diagram of the SACM.

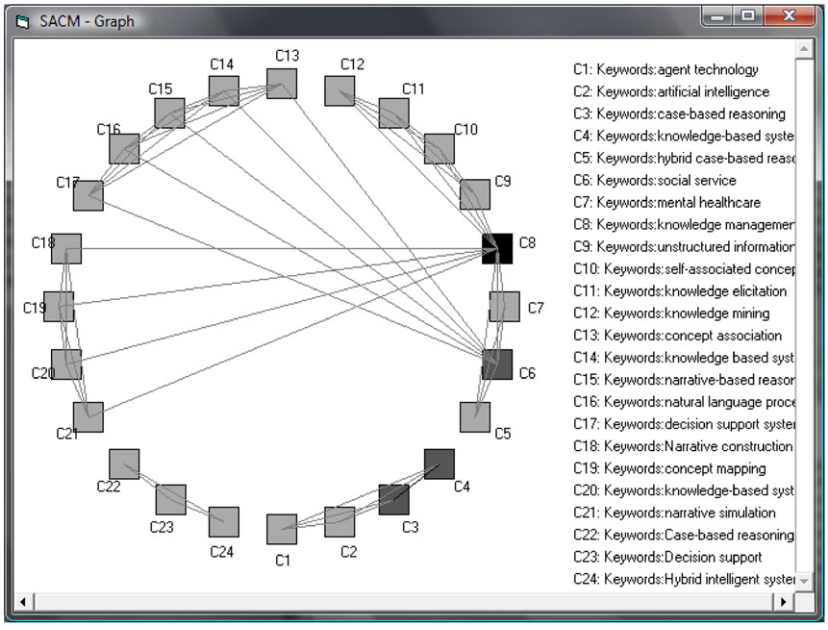

Fig. 3. An example of SACM.

– Assign the degree of importance of relation L_{ij} for each pair of concepts (C_i,C_j), where $i,j \in n$ and n is the total number of distinct concepts **C**, by the following equation:

$$L_{ij} = \text{Min}(W_i,W_j) \quad (2)$$

– If the indexed text is the first record, the temporary SACM is saved as the original SACM and the process goes to Step 1. Otherwise, the process goes to Step 2.

(ii) *Step 2*: combining the temporary SACM with the original SACM

– The nodes and relations of the temporary SACM are matched with that of the original SACM. If any concepts are missing, the weight of that concept and the degrees of importance of that concept's associated relations are assigned to be 0
– The weight of the concepts $\boldsymbol{W}$ and the degrees of relation $\boldsymbol{L}$ of the original SACM are adjusted based on $\boldsymbol{W}$ and $\boldsymbol{L}$ of the temporary SACM by the following equations:

Let $\boldsymbol{W_i}$, $\boldsymbol{W_i'}$, $\boldsymbol{W_i''}$ be the original, temporary and combined weights of C_i, N be the total number of records of the original SACM.

$$\text{for } N>0,\ W_i'' = \frac{W_i' + NW_iW_{\mathbf{max}}}{N+1} \quad (3)$$

$$\text{for } N=0,\ W_i'' = W_i' \quad (4)$$

where

$$W_{\mathbf{max}} = \text{Max}(W_1,W_2,\ldots,W_n) \quad (5)$$

Let L_{ij}, L_{ij}', and L_{ij}'' be the original, temporary, and combined degree of relations respectively, between C_i and C_j, N be the total number of records of the original SACM.

$$\text{for, } N>0,\ L_i'' = \frac{L_i' + NL_iL_{\mathbf{max}}}{N+1} \quad (6)$$

$$\text{for } N=0,\ L_i'' = L_i' \quad (7)$$

$$L_{\mathbf{max}} = \text{Max}(L_1,L_2,\ldots,L_m) \quad (8)$$

(iii) *Step 3*: the parameters of the combined SACM is adjusted.

– The parameters of the combined SACM, $\mathbf{P} = (W_{\mathbf{max}},L_{\mathbf{max}},N)$ is revised by Eqs. (5) and (8).
– The total number of records N is increased by 1.

3.3. Invention support

The invention support process is accomplished by query agent and retrieval agent. They interact with the system users. The invention support process starts with entering the query of the invention problem or concepts by user. Then, the query agent extracts the key concepts from the query and looks up the related concepts based on the extracted concepts to invoke the retrieval agent to retrieve the patent documents that contain those related concepts. The classification agent and relationship agent are also invoked to depict the patterns among the retrieved documents. With the use of the classification agent, retrieved documents can be divided into different preset groups. Based on the SACM analysis, association between concepts can be depicted. Patent maps can be drawn for showing the important concepts and relationships among the patent documents. A snapshot of the patent map generated by the system is shown in Fig. 5.

4. Experiments and results

A prototype of a Semantic-based Intellectual Property Management System (SIPMS) has been established using Hypertext Preprocessor (PHP), Microsoft Visual Basic and Sun Java. A series of experiments have been carried out for comparing the results between the classifications using Latent Semantic Indexing (LSI) and the proposed concept extraction indexing method. LSI is an indexing and retrieval method to identify patterns in the relationships between the terms contained in an unstructured collection of text (Deerwester et al., 1990). It is an instance for comparing similarities between terms from documents using Singular Value Decomposition (SVD) and bag-of-words representation of text documents for detecting words with similar meanings (Faure and Nedellec, 1998; Bisson et al., 2000).

For the testing data, the authors have used the abstracts of the patent documents collected from the United States Patent and

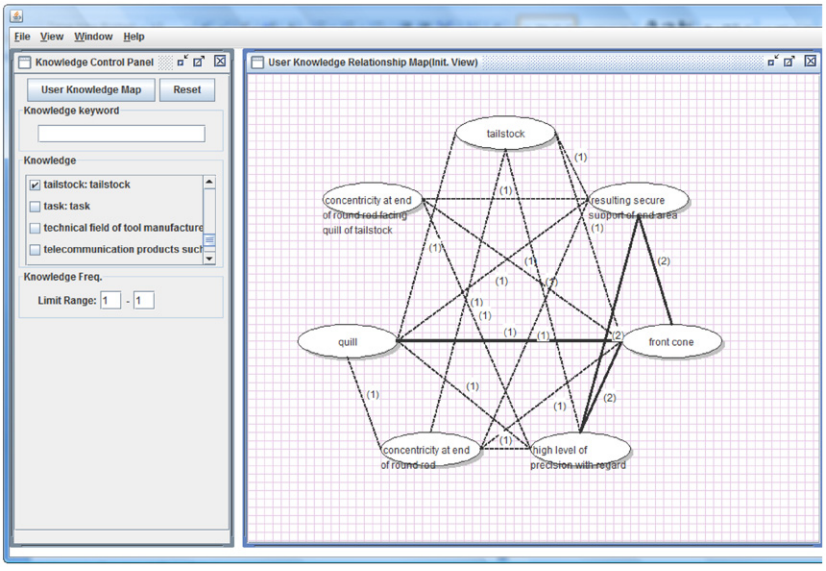

Fig. 5. A snapshot of the Patent map.

Trademark Office (USPTO, http://www.uspto.gov). They are well known databases and the quality of the text is highly assured. Moreover, the main content of the patents are summarized in the abstract and they are highly referred by the knowledge workers. Three different categories (i.e. B32B3804, G06F01900 and H02G1100) are selected randomly from the database. The classes were classified based on international classification.

In the first experiment, 20 abstracts of patents are selected randomly from each of the classes for the experiments (i.e. 60 abstracts of patents in total). Some examples are shown in Appendix C. After entering the texts of the abstracts of the patents, the concepts are automatically extracted. For example, a section of input text drawn from an US patent (US698096) as follows:

> "An earth-moving vehicle includes a carriage, which can move on a terrain in mutually opposite directions of movement along a longitudinal axis of the vehicle, and a station for driving and maneuvering, which is able to rotate with respect to the carriage about a vertical axis, houses a plurality of controls, and is oriented in a direction in which an operator is facing during maneuvering of the controls. The vehicle is further provided with a signaling device and with a control assembly, which activates the signaling device to warn, outside the station, when a direction of movement is actuated that is opposite to the direction of orientation of the station. The control assembly operates automatically according to the relative angular position of the station with respect to the carriage about the vertical axis and according to the command issued for the direction of movement." (Patent No: US6980896).

Table 1
Extracted concepts.

ID	Concepts	Phrase type
1	Earth-moving vehicle	Noun phrase
2	Include	Verb phrase
3	Carriage	Noun phrase
4	can move	Verb phrase
5	Terrain	Noun phrase
6	Opposite direction of movement longitudinal axis of vehicle	Noun phrase
7	Station driving	Noun phrase
8	Maneuvering	Noun phrase
9	Be	Verb phrase
10	Rotate	Verb phrase
11	Respect	Noun phrase
12	carriage vertical axis	Noun phrase
13	House	Verb phrase
14	Plurality of control	Noun phrase
15	Be orient	Verb phrase
16	Direction	Noun phrase
17	Operator	Noun phrase
18	Be face	Verb phrase
19	Maneuvering of control	Noun phrase
20	Vehicle	Noun phrase
21	Be further	Verb phrase
22	Signal device	Noun phrase
23	Control assembly	Noun phrase
24	Activate	Verb phrase
25	Warn	Verb phrase
26	Station	Noun phrase
27	Direction of movement	Noun phrase
28	Direction of orientation of station	Noun phrase
29	Operate automatically	Verb phrase
30	Relative angular position of station respect	Noun phrase
31	Command issue direction of movement	Noun phrase

With the use of the system, the extracted concepts are shown in Table 1. *K*-fold cross-validation is used for the experiments (McLachlan et al., 2004). In *K*-fold cross-validation, the original sample is divided into *K* subsamples. A single subsample is retained as the validation data for testing the model, and the remaining subsamples are used as training data. After indexing by the proposed and traditional method, the documents are then classified by Naive Bayesian algorithm and they are compared with the original international classes. The process is then repeated *K* times, with each of the *K* subsamples used once as the validation data. The results are then averaged to produce a single estimation. In the present study, 30-fold, 20-fold, 15-fold, 12-fold, 10-fold, 6-fold, 5-fold, and 1-fold cross-validations are used. The results are shown in Table 2. From the results, it is interesting to note that the proposed method has outperformed the traditional method in all the validations. The averaged accuracy of classification improves from 5% to 10%.

Another experiment was carried out for measuring the scalability of the algorithm. 300 abstracts of the patents from the three categories (i.e. B32B3804, G06F01900, and H02G1100) were extracted from the USPTO database. 150 abstracts were used as the testing data, and the other 150 abstracts were used incrementally (with a 6 abstracts increment) as the training data. The experimental results are shown in Fig. 6. It is found that the

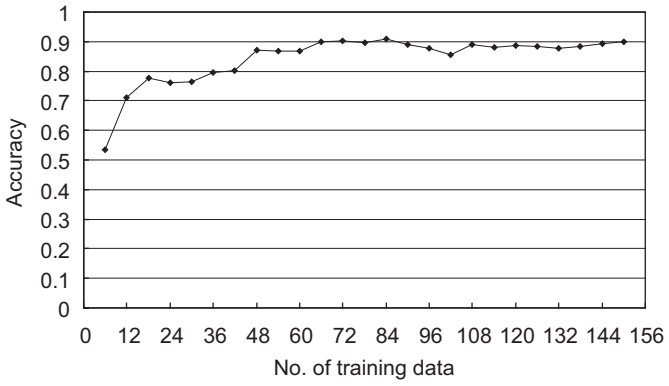

Fig. 6. Experimental results.

Table 2
Results of the experiments.

	Traditional method (LSI)			Proposed method		
	No. of correct classification	No. of incorrect classification	Accuracy (%)	No. of correct classification	No. of incorrect classification	Accuracy (%)
30-fold	50	10	83.3	53	7	88.3
20-fold	50	10	83.3	51	9	85
15-fold	51	9	85	53	7	88.3
12-fold	51	9	85	51	9	85
10-fold	49	11	81.7	55	5	91.7
6-fold	49	11	81.7	51	9	85
5-fold	52	8	86.7	53	7	88.3
1-fold	51	9	85	52	8	86.7

accuracy of the proposed method increases steadily with the increasing number of training data. And the accuracy starts to converge at around 48 training data (i.e. about 1/3 of the amount the testing data). Therefore, it shows that it is capable of processing a vast amount of data when the amount of training data increases.

5. Trial implementation

The methodology in Section 4 has been verified through a trial implementation at a selected reference site. The reference site is a RFID consultancy organization in Hong Kong. A prototype of the SIPMS has been built for supporting its IP management and patent analysis. The organization was facing similar problems found in the industry, such as globalization, increasing competition, and market and product information overload. It is necessary to develop an IPMS to improve IP sharing and management.

Through a study of the document and conducting interviews, the user requirements are collected after a study of the workflow of the company and collecting the required information such as the information flow of the company, the IP management process, the company expectations, etc. Then the project idea is proposed to the management level in order to obtain the management buy-in. This allows systematic handling of IP information and supports the technology invention process of the knowledge workers. After studying the workflow of the organization, a prototype system has been built for the implementation of the SIPMS. The functionalities of the system are designed and shown in Fig. 7 and a snapshot of the system is shown in Fig. 8.

The system provides the basic administrative activities of IP which includes the documentations, maintenance management and data trend analysis. It provides a clear review to let the organizations know what they owned. It also facilitates the internal information flow, store IP assets information, and retains knowledge. The knowledge workers of the organization are invited to provide users' feedbacks of the usage of the system through survey questionnaires and interviews. The questionnaire is designed for evaluating the system usefulness and effectiveness. For system usefulness, there are 13, 9, and 4 evaluation items related to IP portfolio, IP search, and customization, respectively. For system effectiveness, there are 8, 7, and 7 evaluation items related to knowledge access, knowledge use, and knowledge capture, respectively. The details of questionnaires can be found in Appendix B.

Table 3 shows the results of the questionnaire evaluation. From the results, it is interesting to note that the users agree that the system can improve their work in different dimensions discussed above.

6. Conclusion and future work

Effective management of IP is essential for an organization to discover valuable knowledge hidden underneath the sea of information and knowledge. However, the conventional way of IP management which relies on the input of human experts is inadequate. This paper presents SIMPS which serves the purpose of incorporating semantic analysis and text mining techniques for processing and analyzing the unstructured patent documents.

The SIMPS allows for retrieving, automatically classifying, capturing and sharing right knowledge from massive unstructured text under multiple concepts at different levels of abstraction. With the successful development and trial implementation of the SIMPS, the accuracy of categorization of unstructured information is improved. It also allows an organization to capture the valuable knowledge embedded in the patent documents. This helps an organization to explore business opportunities for continuous business improvement. Comparing to the existing technologies, the proposed method extracts the key concepts of the patent documents and discovers the relations among the concepts based on the syntactic structure of the documents. This approach is capable of extracting concept and relationship from the document itself rather than retrieved from predefined dictionary. Hence, it enables the approach to be applied in any domains without initial knowledge capture.

For future work, the system should be evaluated on larger collections with more candidate users since the results are only preliminary as the number of documents considered is rather small. Furthermore, the current study focuses mainly on the unstructured text. However, lots of valuable information can be discovered from the images of patent documents. Further studies may focus on the analysis of different kind of information.

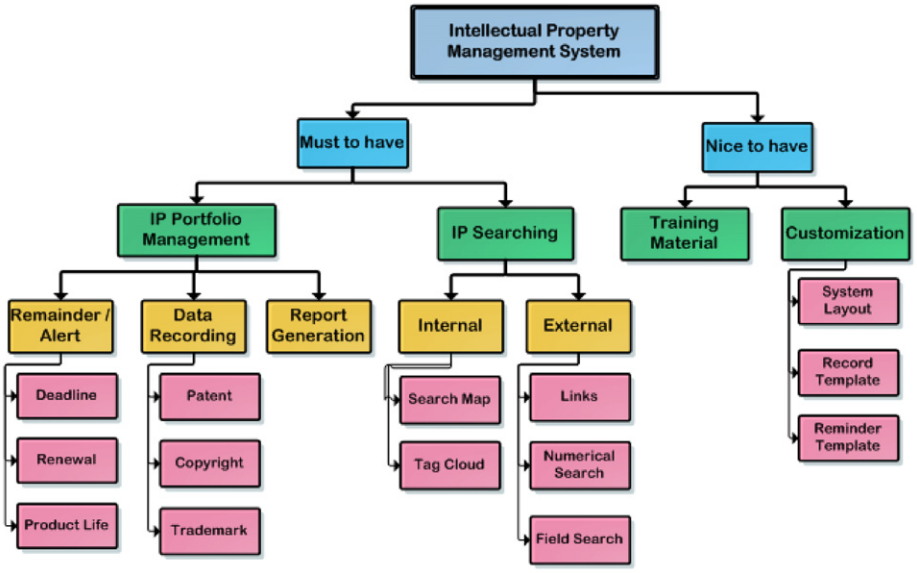

Fig. 7. The functions of the prototype SIPMS.

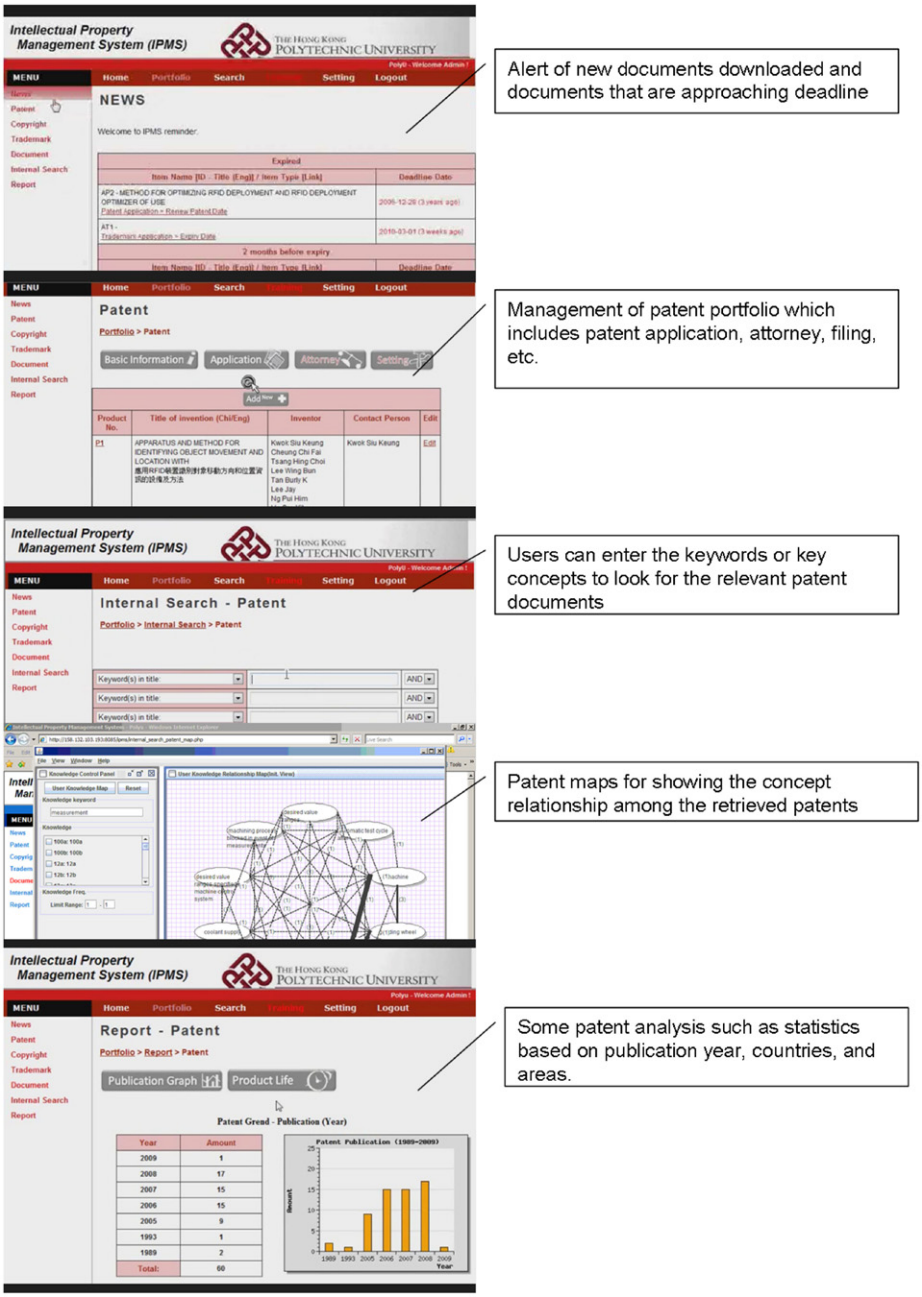

Fig. 8. A snapshot of the SIPMS.

Table 3
Results of the questionnaire evaluation.

Usefulness and effectiveness	Score (%)
Portfolio function	54 out of 65 (83.1)
IP search	38 out of 45 (84.4)
Customization	13 out of 20 (65)
Knowledge access	31 out of 40 (77.5)
Knowledge use	26 out of 35 (74.3)
Knowledge capture	28 out of 35 (80)

Acknowledgments

The authors would like to express their sincere thanks to the Research Committee of The Hong Kong Polytechnic University for financial support of the research work under the Project no. G-YH63.

Appendix A. Tags set

1. cc: coordinating conjunction,
2. cd: cardinal number,
3. det: determiner,
4. ex: existential there,
5. fw: foreign word,
6. in: preposition or subordinating conjunction,
7. jj: adjective,
8. jjr: adjective, comparative,

9. jjs: adjective, superlative,
10. ls: list item marker,
11. md: modal,
12. nn: noun, singular or mass,
13. nns: noun, plural,
14. nnp: proper noun, singular,
15. nnps: proper noun, plural,
16. pdt: predeterminer,
17. pos: possessive ending,
18. prp: personal pronoun,
19. prp$: possessive pronoun,
20. rb: adverb,

Table B1

Characteristic and functions IP portfolio—patent	Not useful 1	Less useful 2	Neutral 3	Useful 4	Very useful 5
1. IP data record—basic information					
2. IP data record—application					
3. Source patent (automatic link with basic information)					
4. Deadline remainder/alert					
5. Report Generator—publication graph					
6. Report Generator—product life					
IP portfolio—copyright					
7. IP data record					
8. Report generator					
IP portfolio—trademark					
9. IP data record—basic information					
10. IP data record—application					
11. Source trademark (automatic link with basic information)					
12. Deadline remainder/alert					
13. Report generator					
IP searching					
14. Internal searching (patent)					
15. Internal searching (copyright)					
16. Internal searching (trademark)					
17. Internal searching—tag cloud					
18. Internal searching—search map					
19. External search—numerical search					
20. External search—field search					
21. Search result exporting (excel)					
22. External links					
Custom setting and additional service					
23. Custom system layout					
24. Custom deadline template					
25. Custom data record page					
26. Training material and sessions					

Comments (optional): e.g. any other additional function(s) can help your company to operate IP assets more effective and efficient?
E.g. any areas need to be improvement on this IPM system?

Table B2

System effectiveness	Strongly disagree 1	Disagree 2	Neutral 3	Agree 4	Strongly agree 5
1. Easy to capture the information you wanted					
2. Easy to access or read the information you wanted					
3. Easy to use the found data to make the decision					
4. The IPM system can facilitate the information retain processes					
5. Easy to understand how to enter data					
6. I can find out the needed/relevant information using IPM system					
7. The information in IPM system can help on managing IP assets					
8. Sufficient instruction on entering data					
9. The "Application" page has clear stage when add the information					
10. All needed data is entry into the system					
11. The data on the system is enough to support our company operation					
12. The table is clear to show the data detail					
13. The IP item (Portfolio) in IPM system cover types of company IP assets					
14. Deadline reminders can give me an alert on perform a pre-action					
15. The deadline templates can increase efficiency of data inputting					
16. It is easier to share patent knowledge through deadline templates					
17. Custom data record can facilitate me on accessing IP data					
18. The "report" function can show the trend and help for decision making					
19. The "internal search" function can search what I want to find					
20. The "external search" is more coefficient for me when search the external database					
21. The exporting function of search result is easier to keep record					
22. The links on the "Search" is more coefficient for me when search the external database					
23. The IPM system can integrate company's Intellectual Property/assets Management					
24. I would like to use this IPM system to record the IP data					

21. rbr: adverb, comparative,
22. rbs: adverb, superlative,
23. rp: particle,
24. sym: symbol,
25. to: to,
26. uh: interjection,
27. vb: verb, base form,
28. vbd: verb, past tense,
29. vbg: verb, gerund or present participle,
30. vbn: with the indexed records. Articiple,
31. vbp: verb, non-3rd person singular present,
32. vbz: verb, 3rd person singular present,
33. wst: wh-determiner,
34. wp: wh-pronoun,
35. wp$: possessive wh-pronoun, and
36. wrb: wh-adverb.

Table C1

Title	Patent ID	International class
Method to create 3-dimensional images from a 2-dimensional image	US7682476	B32B3804
Method for manufacturing circuit forming substrate	US7685707	B32B3804
Phase-changing sacrificial materials for manufacture of high-performance polymeric capillary microchips	US7686907	B32B3804
Reel-type package of flexible printed circuit boards and method for supplying thereof	US7686909	B32B3804
Process for making laminated and curved panels	US7686911	B32B3804
Method for bonding substrates and method for irradiating particle beam to be utilized therefor	US7686912	B32B3804
Method of attaching a label to a thermoplastic substrate	US7691218	B32B3804
Method of depositing functional films on substrates such as glass sheets and film-coating machine for implementing said method	US7691220	B32B3804
Apparatus and method for making fiber reinforced sheet molding compound	US7691223	B32B3804
Method of making a label sheet	US7695584	B32B3804
Rebonding a metallized fabric to an underlying layer	US7695585	B32B3804
Photosensitive epoxy resin adhesive composition and use thereof	US7695586	B32B3804
Method of fabricating a security tag in an integrated surface processing system	US7704346	B32B3804
Method for processing a semiconductor wafer	US7708855	B32B3804
Hot pressing ceramic distortion control	US7708856	B32B3804
Process for making disposable wearing article	US7708857	B32B3804
Tape cutter device	US7712505	B32B3804
Method and device for applying a flat material web section	US7713371	B32B3804
Self-passivating plasma resistant material for joining chamber components	US7718029	B32B3804
Flexible material	USRE41346	B32B3804
Conduit bending system	US6980880	G06F01900
Controller for wire electric discharge machine	US6980879	G06F01900
Earth-moving vehicle including pivotable maneuvering station	US6980896	G06F01900
Electronic control system for agricultural vehicle	US6980895	G06F01900
Estimation of intake gas temperature in internal combustion engine	US6980902	G06F01900
Extensions to dynamically configurable process for diagnosing faults in rotating machines	US6980910	G06F01900
Failure diagnosis apparatus for temperature sensor	US6980904	G06F01900
Information processing apparatus and method, program storage medium, and program	US20060015216	G06F01900
Intergrated reservoir optimization	US6980940	G06F01900
Mechanism for graphical test exclusion	US6980916	G06F01900
Membrane valve controller of a dust-collecting device	US6980878	G06F01900
Method for determining an estimate of the mass of a motor vehicle	US6980900	G06F01900
Method for evaluating remaining electric charge of a battery, and associated single chip system	US20090287434	G06F01900
Phase noise compensation for spectral measurements	US6980915	G06F01900
Phase recovery filtering techniques for SCP throughput shortage	US6980893	G06F01900
Software for improving the accuracy of machines	US6980881	G06F01900
System and method for detecting an anomalous condition in a multi-step process	US6980874	G06F01900
System and method for real-time fault detection, classification, and correction in a semiconductor manufacturing environment	US6980873	G06F01900
Temperature-sensing wafer position detection system and method	US6980876	G06F01900
Walking condition determining device and method	US6980919	G06F01900
Articulating cable chain assembly	US7552581	H02G 11/00
Cable manager for network rack	USRE41353	H02G 11/00
Cable or the like protection and guide device	US7526910	H02G 11/00
Cable protection and guide device	US7677024	H02G 11/00
Cable winding device with clocked keycap and revolving electrical switch	US7017721	H02G 11/00
Closed cable drag chain	US6966527	H02G 11/00
Cord reel adapting device and method	US7044278	H02G 11/00
Cord retainer	US7108544	H02G 11/00
Deformable structure and cable support system	US7484351	H02G 11/00
Electrical feeding unit and cable holder	US7686380	H02G 11/00
Energy guiding device with reduced friction forces	US7310935	H02G 11/00
Filter housing assembly	US7588614	H02G 11/00
Guide chain	US6966434	H02G 11/00
Method of and system for starting engine-driven power equipment	US7148580	H02G 11/00
Power cord cooling apparatus for a vacuum cleaner	US7690485	H02G 11/00
REEL AND REEL HOUSING	US20090057085	H02G 11/00
Retractable cable assemblies and devices including the same	US7032728	H02G 11/00
Retractable electric cord receiving device and ventilation apparatus	US7490706	H02G 11/00
Retractor apparatus for electrical cord	US7527132	H02G 11/00
Skate unit for cable protection and guide device	US7373770	H02G 11/00

Appendix B. Questionnaires for system usefulness and effectiveness evaluation

Part I

Table B1 shows the functions of Intellectual Property Management (IPM) system. Rate the usefulness from 1 to 5 and give comments about the system.

Part II

Table B2 shows the sentences about the functions, characteristics and effectiveness of Intellectual Property Management (IPM) system. Please rate the following statement from 1 (strongly disagree) to 5 (strongly agree)

Appendix C. Patents used for the experiments

See Table C1 for more details.

References

Alves, A., Pereira, F., Cardoso, A., 2002. Automatic Reading and Learning from Text. Proceedings of the International Symposium on Artificial Intelligence.

Archibugi, D., Pianta, M., 1996. Measuring technological change through patents and innovation surveys. Technovation 16 (9), 146–451.

Basberg, B., 1987. Patents and the measurement of technological change: a survey of the literature. Research Policy 16, 131–141.

Be'de'carrax, C., Huot, C., 1994. A new methodology for systematic exploitation of technology databases. Information Processing and Management 30 (3), 407–418.

Bisson, Gilles, Nedellec, Claire, Cañamero, Dolores, 2000. Designing clustering methods for ontology building—the Mo'K workbench. ECAI Workshop on Ontology Learning 2000.

Braam, R., Moed, H.A.A.V.R., 1991. Mapping science by combined co-citation and word analysis. 2. Dynamical aspects. Journal of the American Society for Information Science 42 (2), 252–266.

Brockhoff, K., Chakrabarti, A.K., Hauschildt, J., 1999. Evaluation of dynamic technological developments by means of patent data. The Dynamics of Innovation: Strategic and Managerial Implications 1999, 107–132.

Callon, T., Courtial, J., Laville, F., 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. Scientometrics 22 (1), 155–205.

Campbell, R.S., 1983. Patent trends as a technological forecasting tool. World Patent Information 5 (3), 137–143.

Chakrabarti, A.K., Dror, I., Eakabuse, N., 1993. Interorganizational transfer of knowledge: an analysis of patent citations of a defense firm. IEEE Transactions on Engineering Management EM-40 (1), 91–94.

Cutting, D., Karger, D., Pedersen, J., Tukey, O., 1992. Scatter/gather: a cluster-based approach to browsing large document collections. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval. ACM, 1992, pp. 318–329.

Deerwester, S., Dumais, S., Furnas, G., Landuer, T., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American Society of Information Science 41 (6), 391–407.

Eisen, M., Spellman, P., Brown, P., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. In: Proceedings of the National Academy of Science, USA, vol. 95, pp. 14863–14868.

Ernst, H., 1997. Use of patent data for technological forecasting: the diffusion of CNC-technology in the machine tool industry. Small Business Economics 9, 361–381.

European Patent Office, 2010. 〈http://www.epo.org〉.

Evenson, R., Puttnam, J., 1988. The Yale–Canada Patent Flow Concordance. Economic Growth Center, Yale University, New Haven, CT.

Fall, C.J., Torcsvki, A., Benzineb, K., Karetka, G., 2003. Automated categorization in the international patent dassification. ACM SI IR Forum, 37.

Fattori, M., Pedrazzi, G., Turra, R., 2003. Text mining applied to patent mapping: a practical business case. World Patent Information 25, 335–342.

Faure, D., Nedellec, C., 1998. A corpus-based conceptual clustering method for verb frames and ontology. In: Velardi, P. (Ed.), Proceedings of the LREC Workshop on Adapting lexical and Corpus Resources to Sublanguages and Applications, pp. 5–12.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurasamy, R., 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press.

Grandstrand, O., 1999. The Economics and Management of Intellectual Property: Toward Intellectual Capitalis. Edward Elgar, UK.

Grilliches, Z., 1990. Patent statistics as economic indicators: a survey. Journal of Economic Literature 28, 1661–1707.

Grobelnik, M., Mladenic, D., Jermol, M., 2002. Exploiting text mining in publishing and education. In: Proceedings of the ICML-2002 Workshop on Data Mining Lessons Learned, pp. 34–39.

Hirschey, M., Richardson, V., 2001. Valuation effects of patent quality: a comparison for Japanese and US firms. Pacific-Basin Finance Journal 9, 65–82.

Holl, B., Jaffe, A., Trajtenberg, M., 2000. Market Value and Patent Citations: A First Look. NBER Working Paper Series, Cambridge, MA.

Hufker, T., Alpert, F., 1994. Patents: a managerial perspective. Journal of Product and Brand Management 3 (4), 44–54.

Jain, A.K., Murthy, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Computing Reviews 31 (3), 264–323.

Jung, S., 2003. Importance of using patent information. In: WIPO—Most Intermediate Training Course on Practical Intellectual Property Issues in Business, Organized by the World Intellectual Property Organization (WIPO), Geneva. November 10–14.

Karki, M.M.S., 1997. Patent citation analysis: a policy analysis tool. World Patent Information, 269–272.

Karypis, G., Han, E., Kumar, V., 1994. Chameleon: a hierarchical clustering algorithm using dynamic modeling. IEEE Computer: Special Issue on Data Analysis and Mining 32 (8), 68–75.

Kohonen, T., 1997. Self-Organizing Maps. Springer-Verlag New York, Inc., Secaucus, NJ.

Kruskal, J.B., 1977. Multidimensional scaling and other methods for discovering structure. In: Enslein, K., Ralston, A., Wilf, H.S. (Eds.), Statistical Methods for Digital Computers. Wiley, New York, pp. 296–339.

Lai, Kuei-Kuei, Wu, Shiao-Jun, 2005. Using the patent co-citation approach to establish a new patent classification system. Information Processing and Management 41 (2005), 313–330.

Larkey, Leah S., 1999. A patent search and classification system in digital libraries 99. In: Proceedings of the Fourth ACM Conference on Digital Libraries Merkeley, CA. August 11–14 1999. ACM Press, pp. 79–87.

Lent, B., Agrawal, R., Srikant, R., 1997. Discovering trends in text databases. In: Proceedings of International Conference on Knowledge Discovery and Data Mining, Newport Beach, California, USA. August 14–17.

McLachlan, G.J., Ambroise, K.A., Do, C., 2004. Analyzing Microarray Gene Expression Data. Wiley.

Milic-Frayling, N., 2005. Text processing and information retrieval. In: Zanasi, A. (Ed.), Text Mining and its Applications to Intelligence, CRM and Knowledge Management. WIT Press, Southampton Boston, pp. 1–45.

Miller, A.G., 1995. WordNet: a lexical database for English. Communications of the ACM 38 (11), 39–41.

Mogee, M., 1991. Using patent data for technology analysis and planning. Research-Technology Management 34, 43–49.

Mozina, M., Demsar, J., Kattan, M., Zupan, B., 2004. Nomograms for visualization of naive Bayesian classifier. In: Proceedings of PKDD-2004, pp. 337–348.

Narin, F., Carpenter, M.P., Woolf, P., 1984. Technological performance assessments based on patents and patent citations. IEEE Transactions on Engineering Management, EM-31 (4), 172–183.

Narin, F., Noma, E., 1987. Patents as indicators of corporate technological strength. Research Policy 16, 143–155.

Paci, R., Sassu, A., Usai, S., 1997. International patenting and national technological specialization. Technovation 17 (1), 25–38.

Paik, W., Liddy, E.D., Liddy, J.H., Niles, I.H., Allen, E.E., 2001. Information extraction system and method using Concept-Relation-Concept (CRC) triples. US Patent 6,263,335, July 2001.

Price, L., Thelwall, M., 2005. The clustering power of low frequency words in academic webs. Journal of the American Society for Information Science and Technology 56 (8), 883–888.

Rajaraman, K., Tan, Ah-Hwee, 2002. Knowledge discovery from texts: a concept frame graph approach. In: Proceedings of the 11th International Conference on Information and Knowledge Management, pp. 669–671.

SanJuan, Eric, Ibekwe-SanJuan, Fidelia, 2006. Text mining without document context. Information Processing and Management 42 (6), 1532–1552 December 2006.

Scherer, F., 1982. Inter-industry technology flows in the United States. Research Policy 11, 227–245.

Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, pp. 44–49.

Soo, Von-Wun, Lin, Szu-Yin, Shih-Yao, Yang, Shih-Neng, Lin, Shian-Luen, Cheng, 2005. A cooperative multi-agent platform for invention based on ontology and patent document analysis. In: Proceedings of the Ninth International Conference on Computer Supported Cooperative Work in Design, vol. 1, pp. 411–416.

Tsakalidis, Athanasios, Markellos, Konstantinos, Perdihri, Katerina, Markel Lou, Penelope, Sirmakessis, Spiros, Mayritsakis, George, 2002. Knowledge Discovery in Patent Databases. In: Proceedings of the 11th ACM Conference on Information and Knowledge Management (ACM-CIKM 2002). November 4–9, 2002. McLean VA, US, pp. 672–674.

Velardi, P., Fabriani, P., Missikoff, M., 2001. Using text processing techniques to automatically enrich a domain ontology. In: Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS), pp. 270-284.

Wang, W.M., Cheung, C.F., Lee, W.B., Kwok, S.K., 2008. Self-associated concept mapping for representation, elicitation and inference of knowledge. Knowledge-Based Systems 21 (1), 52–61.

Yoon, B., Park, Y., 2004. A text-mining-based patent network: analytical tool for high-technology trend. Journal of High Technology Management Research 15, 37–50.

Zitt, M., Bassecoulard, E., 1994. Development of a method for detection and trend analysis of research fronts built by lexical or co-citation analysis. Scientometrics 30 (1), 333–351.