國立交通大學

資訊管理研究所

博　士　論　文

專利探勘之研究：專利分類與專利趨勢變化探勘

A Study of Patent Mining:
Patent Classification and Patent Trend Change Mining

研 究 生： 史孟蓉

指導教授： 劉敦仁 博士

中華民國 九十九 年 六 月

專利探勘之研究：
專利分類與專利趨勢變化探勘

研究生：史孟蓉　　指導教授：劉敦仁

摘要

競爭情報有助於企業經營者判定宏觀環境的利基，在企業決策上扮演著相當關鍵的角色。無論是在企業範疇或是總體層面上分析競爭情報，專利資訊絕對是一種衡量企業競爭能力的重要依據。本論文針對不同的專利管理目標提出兩種不同的方法：1)複合式專利分類以達成專利資訊自動化分類及 2)專利趨勢變化探勘以偵測專利研發行為的變化趨勢。

為能更精確的進行專利文件分類，本研究提出的複合式專利分類方法除了整合傳統之內文式、連結式及銓敘資料式專利分類方法外，還包含本研究所提出之專利網路式分類方法。此專利網路除了包含專利文件外，也涵蓋了由專利文件中取得的各種不同特徵作為節點，節點間的連結關係則得自於專利銓敘資料。專利網路式分類法透過分析專利網路中所有可達節點以計算與欲分類專利文件的相關度，並將相關度高的節點作為專利分類的依據。本研究同時也提出一個改良式的 *k-nearest neighbor* 分類器以作為分類之用。我們以從美國專利局(United States Patent and Trademark Office)所收集的專利文件做為測試資料，以評估本研究所提之專利網路式分類及複合式專利分類方法的效能。實驗結果顯示專利網路式分類及複合式專利分類方法皆優於傳統的專利分類方式，其中複合式專利分類方法也優於專利網路式分類方法。

本研究所提出之專利趨勢變化探勘方法可以在不需要專業知識的情況下找出隱含在專利資料中的趨勢變化，可分為專利收集、專利指標計算、及變化探勘三個步驟。在變化探勘階段，本方法將從不同時間區段專利資訊中挖掘出趨勢(以 rule 呈現)，再比較不同時期的專利趨勢以找出趨勢變化，根據趨勢變化的方式可以分成四個種類並分別計算出變化程度，最後將變化程度多寡排序後提供給管理者做為決策之用。我們將專利趨勢變化探勘方法用於台灣的半導體產業分析上，以找出四種不同層級的專利趨勢:競爭對手的研發行為變化、產業領導者在特定技術領域的研發行為變化、產業領導者研發行為變化及產業特定技術領域的趨勢變化。

關鍵字：專利分類，專利趨勢探勘，專利網路，趨勢變化探勘

A STUDY OF PATENT MINING:

PATENT CLASSIFICATION

AND PATENT TREND CHANGE MINING

Student: Meng-Jung Shih Advisor: Duen-Ren Liu

Institute of Information Management
National Chiao Tung University

ABSTRACT

Before executives or mangers make strategic decisions for an organization, competitive intelligence always plays a critical part on identifying niches within macro environment. For analyzing the competitive intelligence, either within a business scope or on a global view, patent is absolutely the most visible factor on evaluating competence of each participating business. This thesis proposes two approaches for different patent management purpose: the hybrid patent classification approach for automatically classifying patents, and the patent trend change mining approach for detecting technological change trends.

The hybrid patent classification procedure combines classic content-based, citation-based and metadata-based methods, with a novel patent network-based method to perform patent classification. The proposed patent network, which contains various types of nodes that represent different features extracted from patent documents, and the nodes are connected based on the relationship metrics derived from patent metadata. The novel approach analyzes reachable nodes in the patent ontology network to calculate their relevance to query patent, after which it uses the *k*-nearest neighbor classifier to classify query patents. To further improve the approach, it is combined with content-based, citation-based and metadata-based classification methods as the proposed hybrid classification approach. We evaluate the performance of the hybrid approach on a test dataset of patent documents obtained from the United States Patent and Trademark Office (USPTO), and compare it with the performance of the three conventional methods. The results demonstrate that the proposed patent network-based approach outperforms the conventional approaches, and the proposed hybrid classification approach performs better than the patent network-based

approach.

The proposed patent trend change mining (PTCM) approach can identify changes in patent trends without the need for specialist knowledge. The proposed approach consists of steps including patent collection, patent indicator calculation, and change detection. In change detection phase, the approach firstly extract rules between two different time periods, comparing them to determine the trend changes. These trend changes are then classified into four categories of change, evaluated with change degree and ranked by their change degree as the output information to be referred by decision makers. We apply the PTCM approach to Taiwan's semiconductor industry to discover changes in four types of patent trends: the R&D activities of a company, the R&D activities of the industry, company activities in the industry and industry activities generally. The proposed approach generates competitive intelligence to help managers develop appropriate business strategies.

Keywords: Patent classification, patent trend mining, patent ontology network, trend change mining

誌謝

本研究能夠順利完成，首先要感謝我的指導教授劉敦仁老師。在交大的這些年，給予我相當大的研究空間，在適當的時間又能解答我對於研究上的疑惑與困境。沒有他當初的提點與指導，我不會有今日豐碩的研究成果。另外，我要感謝我的口試委員王朝煌老師、李瑞庭老師、陳安斌老師，以及羅濟群老師，在我撰寫博士論文期間，他們所給予的寶貴意見與協助，使我的博士論文能夠有更為嚴謹完善的內容。

博士研究期間，有許多快樂美好的回憶，感謝在這段期間陪伴我的諸位同學朋友跟學弟妹，讓我的生活過得充實且多采多姿，深深感謝大家。

我更要感謝我親愛的父母以及妹妹，他們對我不變的信心以及鼓勵，是我完成學業的最大原動力。還有我的先生昶瑞與女兒，在我論文研究過程中給我最大的支持，讓我擁有最強的後盾。

在完成這份論文的過程中，經歷許多人生歷程的曲折，而今能繳出一份令人滿意的成果，最要感謝的是還是這一路上曾幫助過我的師長與親友們，感謝你們，我願把這份榮耀和你們分享，謝謝！

孟蓉 2010.7

CONTENTS

TABLE

FIGURE

Chapter 1 Introduction

For a competitive organization, competence management is critical to organization development and even to survival issue. Complete competence management generally consists by processes including competence identification, assessment, acquisition and knowledge usage (Berio and Harzallah, 2007). Among the four processes of competence management, for building a solid structure of competence that can establish a business in an unassailable position, the key point is to determine which competence in hand and which competence to obtain. To accomplish this task, competitive organizations need to keep tracing the trends of competence change and find potential elements which may substantially improve the organization competitiveness. Unfortunately, most competences, especially competitive intelligence, are neither structured nor quantifiable. So how to effectively discover the trends of change among these abundant unstructured valuable data like documents of intelligent properties or patents will be very essential to an organization to “lock on” the target competences to obtain.

In practice view, the most representative form of the competitive intelligence of an industry is patent. Patent is one of the most valuable yields developed from an innovative idea and is essential for a business or even for an industry to position their values (Guan & Gao, 2009; Su et al., 2009). Accompanied with rapid development of modern technology, patents were developed with a fast increasing speed for decades and in nowadays have accumulated to a large volume. These patent documents embody technological novelty and serve as important sources of competitive intelligence with which enterprises gain strategic advantages (Stembridge & Corish, 2004). Besides the direct benefits from the context of the patents, the accumulation of patents also provides valuable information for strategic decision making. From the vast amount of patents we may extract strategic information of technological trends and changes happened before or emerging nowadays within an industry, and tactical information for identifying patents to acquire. These information are especially useful when a business conducting important decisions about investments, like founding business units versus mergers and acquisition, on the competitive environment within an industry. How to effectively manage the patent documents and to generate valuable information from these precious intellectual properties is becoming an important issue and directly helps the quality of executive decisions of vast industrial investments.

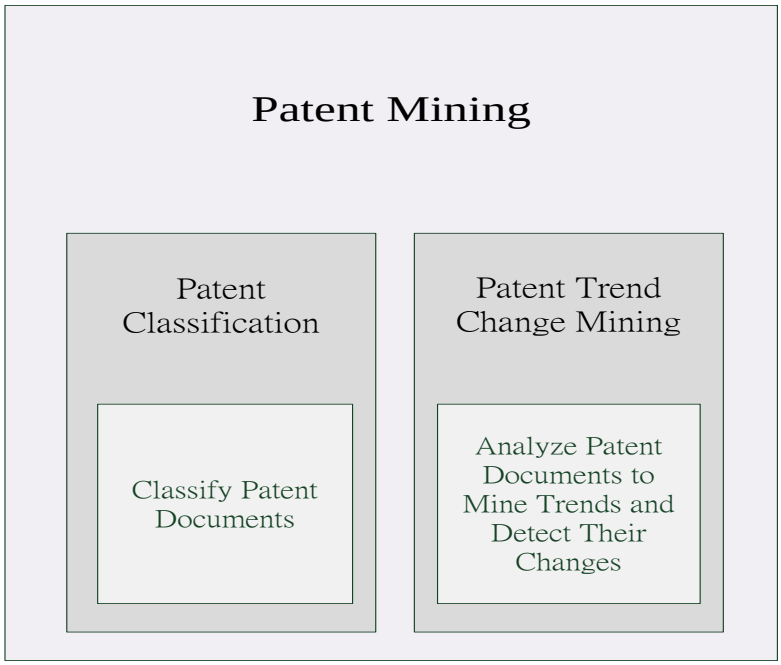

Figure 1. The roles of the proposed approaches in patent mining

In this thesis, two methodologies are proposed for patent mining in separate aspects: a hybrid patent classification approach and a patent trend change mining approach (Shih et al., 2010), as shown in Figure 1. Before introducing the proposed approaches, we discuss their positions and the difficulties respectively.

Many different technological fields have their sets of patents respectively and all the patents belong to specific categories for their practical use. Patent classification is an important step to classify the patent documents. Generally it is a laborious work accomplished by manpower because of the unstructured free-text style of patent context. Patent classification are mostly based on authority-defined classes like United States Patent Classification (UPC) schemes or based on classes defined by business users. In fact many patents are difficult to classify because of their generality or belonging to undefined classes. They may be highly involved in many different fields and also play key roles to competence. Therefore, the complicated relations and connections among different patents shall be considered for industrial analysis. Obviously, there is a pressing need of an effective automatic patent classification approach.

Patent trend change mining is also essential to the patent analysis. Any variation on patent trends in an industry as a whole will directly influence the research and development strategies of all involved enterprises. It emerges when a novel technique developed or when a revolutionary product (or parts) are invented. To maintain a leading position in the highly competitive business environment, enterprise managers need comprehend key intelligence

properties of their own organization, of their competitors, and of the environment in which they operate. By analyzing patent data, managers can evaluate and understand trends in the development of technologies and plan suitable strategies (Stembridge, 2005).

For patent classification part, it covers our proposed novel patent network-based classification approach (see chapter 3) and our proposed hybrid patent classification approach (see chapter 4) that combines classic content-based, citation-based, metadata-based methods and a novel patent network-based method to perform a more effectively patent classification (see chapter 5). For patent trends change mining part, it is comprised of patent trend mining process and change detection process, the details are described in chapter 6.

Chapter 2 Related Works

In this chapter, we present an overview of state-of-the-art patent classification, ontology-based network analysis, association rule mining, and change mining techniques. Then we introduce patent analysis and discuss commonly used patent indictors.

2.1 Patent classification

Patent classification schemes classify patent documents. In recent years, a considerable number of such schemes have been proposed (e.g., Kim & Choi, 2007; Kohonen, et al., 2000; Lai & Wu, 2005; Larkey, 1999; Richter & MacFarlane, 2005; Cong & Tong, 2008; Cong & Loh, 2010; Trappey, et al., 2006). The features extracted from patent documents for classification purposes can be divided into three types: content features, citation information and metadata. The detailed parts of patent documents are showed in Appendix C.

2.1.1 Content-based patent classification

Since patent classification is formulated as a text categorization problem that involves assigning a patent document to the correct class, most studies only consider patent content information to address the problem (e.g., Loh, et al., 2006). In content-based patent classification approaches, the content of a patent document d_p is represented by a vector of term weights, $\overrightarrow{d_p} = \langle w_{1p}, \ldots, w_{|T|p} \rangle$, where T is the set of terms. The similarity of two patent documents is defined as the cosine value of their term vectors (Yang, 1994). The most popular term weighting function is term frequency / inverse document frequency (*tfidf*), developed by Salton and Buckley (Salton & Buckley, 1988). It is defined as follows:

$$tfidf(t_k, d_p) = \#(t_k, d_p) \times \log(N/n_{t_k}), \quad (1)$$

where $\#(t_k, d_p)$ denotes the number of times term t_k occurs in patent document d_p (the term frequency); and $\log(N/n_{t_k})$ represents the total number of patent documents divided by those in which t_k occurs (the inverse document frequency).

The similarity of two patent documents is defined as the cosine value (Yang, 1994) of their respective term vectors, as shown in Eq. 2:

$$Sim(q, p) = \frac{\overrightarrow{d_q} \cdot \overrightarrow{d_p}}{|\overrightarrow{d_q}||\overrightarrow{d_p}|} \quad (2)$$

where q is the query patent document to be classified; and p is a patent document in the training patent dataset.

Based on the similarity of patent documents, the *kNN* classifier selects the k-nearest neighbors of a query patent to predict the class of the patent based on majority vote. The class that most of neighboring patents belong to is chosen as the class of the query patent.

Instead of using the full text of a patent document as the basis for classification, some approaches classify patent documents by considering normative sections, such as the abstract, background, and results (Kim & Choi, 2007; Fall, 2003, 2004; Larkey, 1999; Cong & Tong, 2008; Loh, et al., 2006; Trappey, et al., 2006). These studies regard the patent document's abstract as the most informative feature (Larkey, 1999; Liang, et al., 2003; Loh, et al., 2006).

2.1.2 Citation-based patent classification

In real-world applications, patent documents are linked through citations that imply the connections and relationships between the citer and the cited. Approaches that utilize citations have been proposed (Lai & Wu, 2005; Li et al., 2007). These studies demonstrate that citation-based patent classification performs better than content-based classification. In our work, we also consider the citation relationships between patent documents when constructing the patent ontology network.

2.1.2.1 Co-citation patent classification

The co-citation approach (Lai & Wu, 2005) classifies a query patent according to the majority vote of the classes of its cited patents. For example, suppose a query patent cites five documents in the basic patent set. If three of the cited patents belong to class *C1* and the other two belong to class *C2*, the query patent will be assigned to class *C1*. Note that the co-citation approach uses the grouping result of patents, which are clustered according to the co-citation frequency and linkage strength of each pair of basic patents, as the classes, rather than the well-known UPCs (United States Patent Classification) or IPCs (International Patent Classification).

2.1.2.2 Citation network patent classification

In Li et al.'s (2007) approach, every patent has its own citation network in which each cited node is labeled with its classification class. A patent's class is determined by evaluating the similarities between its citation networks and those of other patents already classified into UPC categories. The network similarity, or graph similarity, of two patents is calculated by comparing their random walk paths. This approach adopts a three-stage kernel-based technique for patent classification: data acquisition and parsing, kernel construction, and classifier training. Li et al. (2007) use support vector machine *(SVM)* as the kernel machine. In their approach, the kernel value, namely the patent similarity of a patent pair is calculated as Eq. 3:

$$K(G_{pi}, G_{pj}) = \sum_h \sum_{h'} l(h, h') O(h|G) O(h'|G'), \quad (3)$$

where G_{pi} and G_{pj} represent the citation networks associated with two patents p_i and p_j; h and h' are the random walk paths in the respective graphs; and $O(h|G)$ and $O(h'|G')$ denote the probability of random walk paths that exist in the citation networks. $l(h|h')$ is defined as follows:

$$l(h|h') = \begin{cases} 1, \text{if } h \text{ and } h' \text{ are identical} \\ 0, \text{otherwise} \end{cases}$$

For each class, the SVM classifier will generate a classification model. The kernel matrix is an augmented matrix which contains patent similarity vectors of all patents in the training set and their respective class labels. The class label of each patent is defined as whether the patent belongs to a specific class—the label is 1 if a patent belongs to the class, and is -1 otherwise. This is so called one-against-rest model for the SVM to handle multiclass problems. For each specific class, its well-trained SVM model can be used to predict if a query patent belongs to the class. The final class is then determined with winner-takes-all strategy from all these SVM models of classes.

2.1.3 Metadata-based patent classification

Metadata is defined as "information that describes data". The metadata in a patent document, such as inventors' names and assignees' names, may be correlated with the document's content and can be used for classification purposes. Richter & MacFarlane (2005) showed that patent classification based on a document's metadata can improve the accuracy of the results. Their approach uses metadata, such as the inventor's name, the applicant's name and the IPC code to help classify commercial intellectual property. Because the approach considers text, inventor and IPC metadata simultaneously, it yields a better classification result. Patent documents are mapped into vectors of terms, inventors' names and IPCs. For the text, the weights of terms are calculated by the *tfidf* approach (Salton and Buckley, 1988); the weight of each inventor is calculated as $\sqrt{1/\#inv}$, where *#inv* is the total number of inventors of the patent; and the weight of each IPC code is calculated as $\sqrt{1/(\#ipc+1)}$, where *#ipc* is the number of IPC code assigned to the patent. Note that the primary IPC is weighted twice as high as other IPC assigned to the patent. After compiling the vectors, the similarity between two patent documents can be calculated. The *kNN* classifier is then used to identify the class of the query patent based on the similarity (cosine value) of patent documents.

One limitation of the above method is that it only works well when the inventors of a query patent also exist in the training set. The method does not utilize indirect relationships to help classify patents developed by new inventors who are not included in the training set. In contrast, our method constructs a patent ontology network; thus, indirect relationships can be used to classify patent documents more flexibly and accurately.

2.2 Ontology-based network analysis

A social network is a social structure made of individuals (or organizations), which are connected by one or more specific types of interdependency, such as friendship, common interest etc. Nodes are the individual actors within the networks, and connections are the relationships between the actors. Social networks have been used to examine how individuals interact with each other, characterizing the many informal connections that link executives together, for example, form community of practice (CoP, i.e., groups of individuals interested in a particular job, procedure, or work domain) or assist knowledge sharing (O'Hara, et al., 2002; Yuan, et al., 2010).

O'Hara, et al. (2002) developed an ontology-based network analysis method to examine ontology-based social networks that help identify CoP. The ontology-based social network is formed by object instances (e.g. person, paper, conference) and semantic relationships (e.g. authorOf, attended) between instances. The rationale behind the method is that the relevance values of nodes increase with the number of semantic paths leading to the object of interest. The instances and their relationships in the ontology network are analyzed by a breadth-first, spreading-activation search algorithm that traverses the semantic relations between instances. In this approach, the relationships and their weights are selected manually and pre-defined.

The purpose of social network analysis is to determine the interactions between a query node (e.g. a person) and nodes (e.g. related persons) in a social network. With a similar concept, we propose to construct an ontology-based patent network for patent class prediction. We modify the ontology-based network analysis method (O'Hara, et al. 2002) and use it for patent network analysis to measure the relevance of a query patent and the nodes in a patent ontology network. The weights of relationships are generated automatically according to the semantic relevance of two nodes. Then, the *k* nodes with the highest relevance to the query patent are used to predict the class of the patent.

2.3 Association rule mining

Data mining techniques have been widely used in various fields of information science (Chang et al., 2009; Kuo, Lin & Shih, 2007; Yen & Lee, 2006; Chen & Liu, 2004; Ngai et al., 2009). Association rule mining is a data mining technique used in various applications, such as market basket analysis. The technique searches for interesting associations or relationships among items in a large data set (Han & Kamber, 2001). Different association rules express different regularities that exist in a dataset; and two measures, support and confidence, are

used to determine whether a mined rule is a regular pattern (Han & Kamber, 2001; Ian & Eibe, 2000). The support measure determines the probability that a transaction contains both the conditional and consequent parts of a rule, while the confidence measure is the conditional probability that a transaction containing the conditional part of a rule also contains the consequent part. The apriori algorithm (Agrawal & Skrikant, 1994) is typically used to find association rules by discovering frequent itemsets (sets of items), which are considered to be frequent if their support exceeds a user-specified minimum support threshold. Association rules that meet a user-specified minimum confidence can then be generated from the frequent itemsets.

In this work, we apply association rule mining to patent data to find patent patterns (rule patterns).

2.4 Change mining

The objective of change mining is to discover changes in two datasets (e.g., about customer behavior) belonging to different time periods. Change mining approaches can be classified as follows:

(a) Decision Tree Models: this method constructs decision trees for two datasets, and then identifies the differences by comparing the two decision trees (Liu et al., 2000; Liu & Hsu, 1996).

(b) Association Rules: this method determines changes by comparing the association rules mined from two datasets (Song, Kim & Kim, 2001; Chen, Chiu & Chang, 2005; Liu, Hsu & Ma, 2001). Users can decide the type of rule changes according to the similarities and differences between the rules in the datasets. There are several types of change mining patterns (Song, Kim & Kim, 2001; Chen, Chiu & Chang, 2005):

- Emerging patterns: the concept of emerging patterns captures significant changes between datasets. An emerging pattern is a rule pattern whose support increases significantly from one dataset to another.

- Unexpected consequent changes: these changes are found in newly discovered association rules whose consequent parts differ from those of the previous rule patterns.

- Unexpected condition changes: these changes are found in a newly discovered association rules whose conditional parts differ from those of previous rule patterns.

- Added rules: these are new rules that only exist in the present dataset.

- Perished rules: these are rules that only exist in the previous dataset.

Association rule change mining techniques are used to analyze transaction data and discover changes in customer behavior. In this work, we identify changes in patent trends from patent data.

2.5 Patent analysis

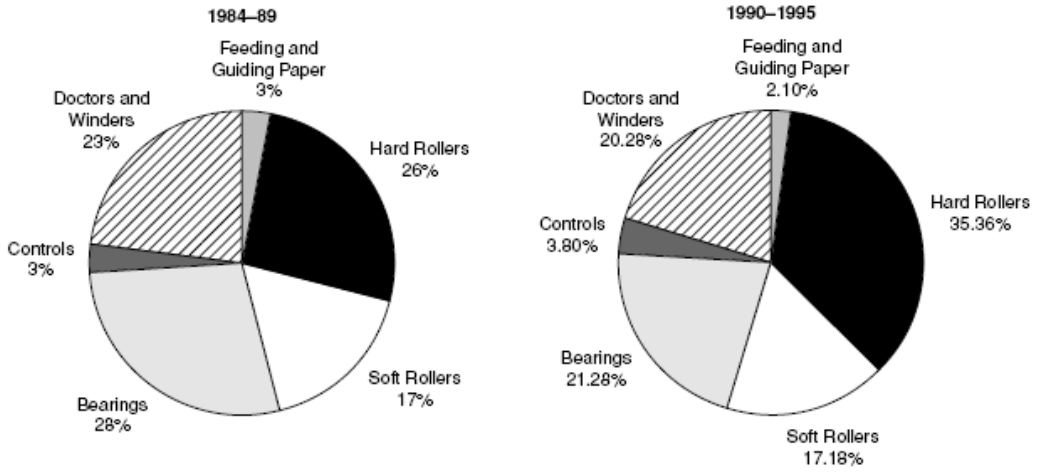

Figure 2. Distribution of technological fields of paper-making machinery

Rapid technological development has made it easier for companies to search and access patent documents. Many patent offices already allow free download of the abstracts and complete texts of their patents (e.g., WIPO (WIPO, 2007), USPTO (USPTO, 2007) and EPO (EPO, 2007)).

Several software tools and services have been developed in the patent field (Breitzman & Mogee, 2002; Dou et al., 2005; Dürsteler, 2007; Huang et al, 2008). These tools analyze patents by classification, clustering, and statistical methods to find the relationships between patents with similar content / structure. The results of patent analysis are usually presented as graphs or tables, and provided to specialists, researchers, and R&D practitioners to help them plan their strategies.

Patent information can be analyzed either quantitatively or qualitatively (Huang et al., 2003). Quantitative measures are based on statistical processing, and indicate the level of patenting activity of an analytical unit (e.g., the number of patents owned by an assignee). Qualitative measures are calculated according to citation information and used to assess the quality of a patent.

In the literature, and in practice, several indicators are used to measure patents quantitatively

or qualitatively. In the next subsection, we introduce patent indicators.

Although existing patent analysis tools can provide various results, analysts still need to compare the results of two periods to identify changes over time. For example, Figure 2 shows the distribution of the technological fields of paper-making machinery in two periods, 1984-1989 and 1990-1995 (Breitzman & Mogee, 2002). Patent analysts can discover changes in the technological field by comparing the two distributions. In this case, R&D activities increased for Hard Rollers and Controls, decreased for Bearings, and remained stable for other areas (Breitzman & Mogee, 2002). Making such comparisons requires professional knowledge. Moreover, changes cannot be ranked intuitively; the degree of change must be calculated and ranked by analysts.

The motivation of this study is to discover changes in the patent trends of different time periods without the need for expert knowledge, and report changes to business managers by ranking the degree of change.

2.6 Patent indicators

Patents are one of the major sources of technological and competitive information because such data is easy to access and the content is highly innovative. Since the value of patents is rarely observable, scholars and research organizations have defined a number of patent indicators to determine the value of patents (Brockoff, 1991; CHI-Research; Reitzig, 2004; Tuomo, Hermans & Kulvik, 2007).

The common patent indicators are described below (Brockoff, 1991; CHI-Research; Reitzig, 2004; Tuomo, Hermans & Kulvik, 2007):

- Patent age: the age of a patent (the patent's age is calculated from the date the patent was applied for).
- Citation made (backward citations): the number of patents cited by the target patent.
- Citation Index (forward citations): the number of citations received by the target patent. It is a measure of the impact of the target patent.
- Originality: the originality of a target patent indicates the diversity of cited patents, i.e., the patents cited by the target patent. The measure is based on the distribution (ratio) of cited patents over classes, as expressed in Eq. 4.

$$Originality = 1 - \sum_{j \in S_B} B_j^{\,2}$$

$$B_j = \frac{\text{Number of cited patents belonging to Class j}}{\text{Number of cited patents}} \quad (4)$$

$$S_B : \text{the set of classes of cited patents}$$

Generality: the generality of a target patent indicates the diversity of citing patents, i.e., the patents that cite the target patent. The measure is based on the distribution (ratio) of citing patents over classes, as expressed in Eq. 5.

$$Generality = 1 - \sum_{j \in S_F} F_j^{\,2}$$

$$F_j = \frac{\text{Number of citing patents belonging to Class } j}{\text{Number of citing patents}} \quad (5)$$

$$S_F = \text{the set of classes of citing patents}$$

- Technology Cycle Time (TCT): the TCT of a target patent is the median age of the patents cited by the target patent. It is a measure of technological progress.

Chapter 3 Patent Network-based Patent Classification

In this section, we introduce the proposed patent network-based classification approach, as shown in Fig. 3. The proposed patent network-based classification approach is implemented in two phases: 1) patent network construction; and 2) patent class prediction phase, which includes patent network analysis, k nearest neighbor extraction and patent class identification.

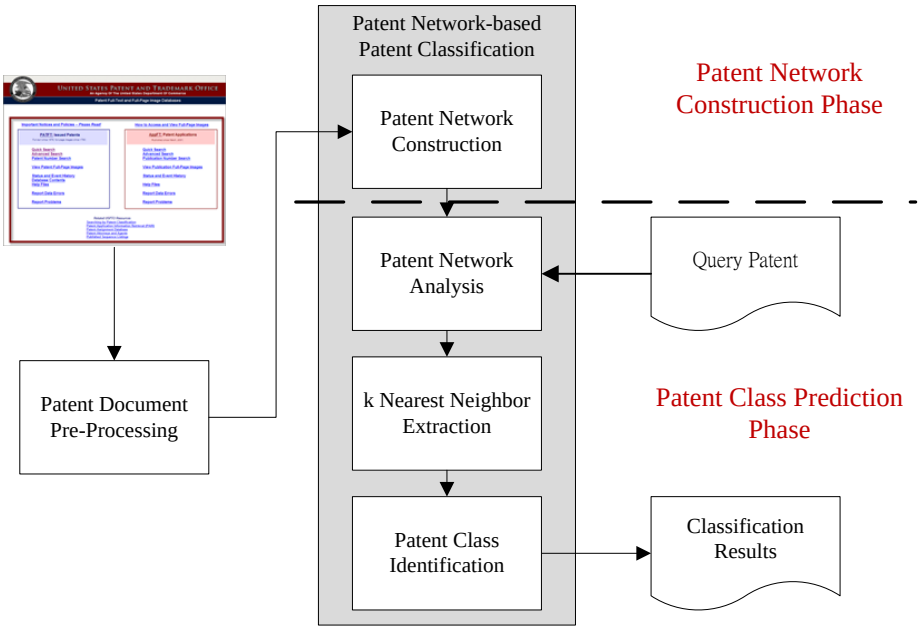

Figure 3. The process of patent network-based patent classification

3.1 Patent Document Pre-processing

In this stage, we first collect patent documents from various sources on the Internet, e.g., the United States Patent and Trademark Office (USPTO). All the patent documents downloaded from USPTO are in HTML format and semi-structured. Therefore, we use a pre-processing module to clean and parse the unstructured texts and transform them into structured data. We also extract the following information from the original documents for further analysis: the patent number, the United States Patent Classification (UPC) code, inventor and assignee names, and citation data.

3.2 Patent Ontology Network Construction

The first step of patent network-based classification process involves building a patent ontology network, as shown in Fig. 4. The relations between instances (nodes) are identified to construct the network. The weights of all the relationships among nodes are derived by the functions described in this section. Relationships (connections) of zero degree are dropped and the network is trimmed to form the final patent ontology network for classification. The proposed patent ontology network contains four types of instances (nodes) and eight types of relations (edges). The node types are patent, UPC class, inventor, and assignee (e.g., a research institute). The weights of the relationships are calculated by the functions listed in Table 1.

$R_{PP}(p_1, p_2)$ denotes the relationship between two patents p_1 and p_2. Both citations and co-citations are considered active relations between two patents, as shown in Eq. 6:

$$\begin{cases} R_{PP}(p_1, p_2) = w_{cite} \times Cite(p_1, p_2) + w_{co\text{-}cite} \times CoCite(p_1, p_2) \\ w_{cite} + w_{co\text{-}cite} = 1 \end{cases}, \quad (6)$$

where $Cite(p_1,p_2)$ is the citation relation between p_1 and p_2 defined as

$$Cite(p_1, p_2) = \begin{cases} 1, \text{if the citation exists (either } p_1 \text{ cites } p_2 \text{ or } p_2 \text{ cites } p_1) \\ 0, \text{otherwise.} \end{cases}$$

and $CoCite(p_1,p_2)$ is the degree of co-citing between p_1 and p_2 defined as

$$CoCite(p_1, p_2) = \frac{|CitedBy(p_1) \cap CitedBy(p_2)|}{|CitedBy(p_1) \cup CitedBy(p_2)|}$$

where $CitedBy(p_1)$ and $CitedBy(p_2)$ are the sets of patents cited by p_1 and p_2, respectively.

$R_{II}(v_1, v_2)$ represents the degree of patents that belong to two inventors v_1 *and* v_2, and is defined as Eq. 7.

$$R_{II}(v_1, v_2) = \frac{|Patents\ (v_1) \cap Patents\ (v_2)|}{|Patents\ (v_1) \cup Patents\ (v_2)|}\ , \tag{7}$$

where *Patents*(v_1) and *Patents*(v_2) are the sets of patents belonging to v_1 and v_2, respectively.

$R_{CI}(v_2, c_1)$ represents the ratio of patents belonging to a specific inventor v_2 to the number of patents in a patent class c_1, and is defined as Eq. 8 :

$$R_{CI}(v_2, c_1) = \frac{|Patents\ (v_2) \cap Patents\ (c_1)|}{|Patents\ (c_1)|}\ , \tag{8}$$

where *Patents*(c_1) is the set of patents belonging to class c_1..

$R_{CA}(c_1, a)$ represents the importance and maturity of a technology of assignee a in a specific technology field, i.e. class c_1 , as shown in Eq. 9:

$$R_{CA}(c_1, a) = \frac{\sum_{p_i \in \text{Patents}\ (a) \cap \text{Patents}\ (c_1)} \text{NumCitations}\ (p_i, a, c_1)}{\sum_{p_j \in \text{Patents}\ (C_1)} \text{NumCitations}\ (p_j, c_1)}\ , \tag{9}$$

where *NumCitations*(p_i, a, c_1) is the number of patents in class c_1 that cite assignee a's patent p_i ; and *NumCitations*(p_j, c_1) is the number of patents in class c_1 that cite patent p_j.

Figure 4 shows an example of a patent ontology network that includes the four types of nodes, i.e., patent, class, inventor and assignee. The weights of relations are calculated using the equations listed in Table 1.

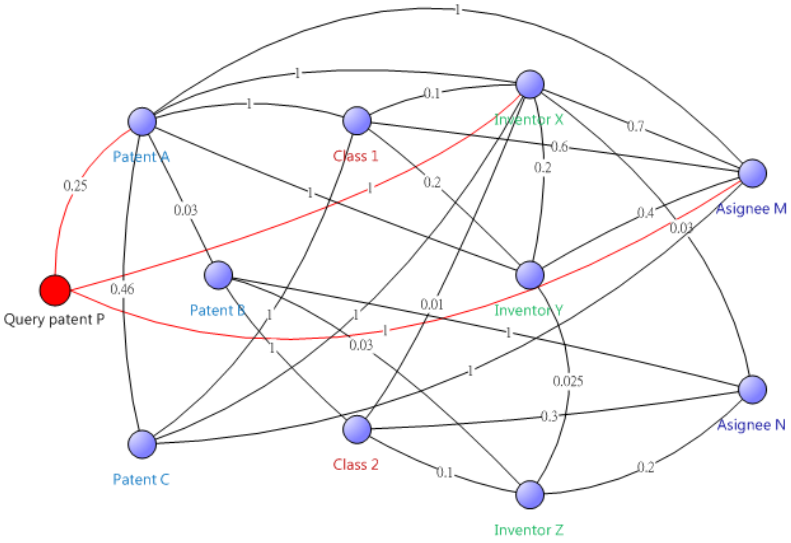

Figure 4. An example of patent ontology network

The patent ontology network is a base map for classifying unclassified patents. In the next sub-section, we describe the classification process based on generate a patent network analysis. Classifying a patent and assigning it to the most suitable class involves three steps: patent network analysis, *k*-nearest neighbor extraction and patent class identification.

Table 1. The relationship metric in the patent ontology network

Relationship Weights	patent p_2	class: c_2	Inventor: v_2	Assignee: a
patent p_1	$R_{PP}(p_1, p_2)$	$R_{PC} = \begin{cases} 1: p_1 \in c_2 \\ 0: p_1 \notin c_2 \end{cases}$	$R_{PV} = \begin{cases} 1: p_1 \text{ invented by } v_2 \\ 0: \text{not related} \end{cases}$	$R_{PA} = \begin{cases} 1: p_1 \text{belonging to } a \\ 0: \text{not related} \end{cases}$
class c_1		N/A	$R_{CI}(v_2, c_1)$	$R_{CA}(c_1, a)$
Inventor: v_1			$R_{II}(v_1, v_2)$	$R_{IA} = \begin{cases} 1: v_1 \text{belonging to } a \\ 0: \text{not related} \end{cases}$

3.3 Patent Network Analysis

To classify a patent document, we first search the patent ontology network to find patent nodes, inventor nodes and assignee nodes that have connections with the query patent. For example, in the network in Figure 4, *X* is the inventor of query patent *P* and the assignee is *M*. Patent *P* also has citation relationships with other patents. These connections are therefore

evaluated to derive their respective weights using the equations listed in Table 1.

After determining all the connections and weights between the query patent and the nodes in the patent ontology network, we calculate the relevance of the query patent to each node in the patent ontology network. The algorithm used for patent network analysis is a modification of the ontology-based network analysis algorithm developed by O'Hara et al. (2002) for identifying an individual's communities of practice. Our algorithm calculates the weights of the nodes and their relations to derive their relevance scores to the query patent. More specifically, it implements a breadth-first, spreading-activation search and traverses the relations between the nodes until it reaches a link threshold, which is the maximum number of consecutive links between nodes that can be traversed. The detailed steps of the patent network analysis algorithm are listed in Appendix A.

3.4 K- Nearest Neighbor Extraction

After calculating the relevance of the query patent document to the nodes in the patent ontology network, the *k* nodes with highest relevance scores to the query patent document are extracted and used to identify the most appropriate class for a patent.

3.5 Patent Class Identification

Let S_q be the set of neighboring nodes identified in the step of *k*-nearest neighbor extraction. In this step, the nodes in *Sq* are used to determine the class of the query patent *q*. Unlike the classical *kNN* method, which can only find neighboring nodes of the same type, the proposed method can find k nodes of various types by using the result of patent network analysis. We only use patent and class nodes to calculate the scores of candidate classes because they are more suitable for interpreting patent classes. For "patent" nodes, the more relevant a patent node *q* is to the query patent, the greater the likelihood that the query patent belongs to the class of that patent node. In addition, for "class" nodes, the more relevant a class node *c* is to the query patent, the greater the likelihood that the query patent belongs to the class of that node. We denote the set of identified neighboring patent nodes and the set of identified neighboring class nodes as S_q^P and S_q^C, respectively. Note that $S_q^P, S_q^C \subset S_q$.

The next step evaluates the predicted scores of candidate classes, which are selected from the identified patent nodes and class nodes. The predicted score $F_{q,c}^{PNW}$ for a given query patent *q* belonging to class *c* is calculated as Eq. 10:

$$F_{q,c}^{PNW} = \sum_{d \in S_q^P} w_d^{PNW} B_{d,c}^P + \sum_{d \in S_q^C} w_d^{PNW} B_{d,c}^C \quad (10)$$

where w_d^{PNW} denotes the weight, namely the relevance score of node d obtained by patent network analysis; and $B_{d,c}^P$ and $B_{d,c}^C$ are defined as follows:

$$B_{d,c}^P = \begin{cases} 1, \text{ if node } d \text{ represents a patent belonging to class } c \\ 0, \quad \text{otherwise} \end{cases}$$

$$B_{d,c}^C = \begin{cases} 1, \text{if node } d \text{ represents class } c \\ 0, \quad \text{otherwise} \end{cases}$$

After obtaining all the predicted scores of classes in C, the class with highest score is taken as the class of the query patent.

Chapter 4 Hybrid Patent Classification

In this section, we propose a hybrid approach that utilizes patent metadata and considers the semantic structure of the patent ontology network. The approach mainly contains two phases, conducting different approaches of patent classification and the combination of class predictions, as shown in Figure 5.

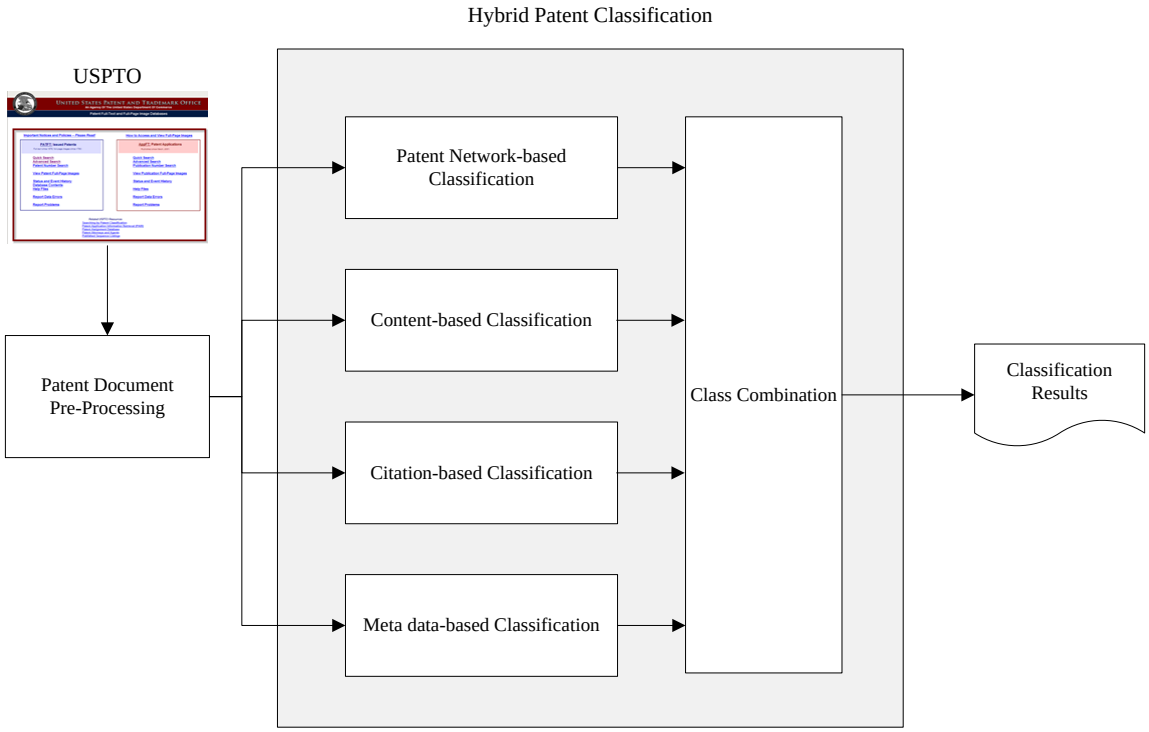

Figure 5. The hybrid patent classification approach

4.1 Patent Classification by Various Methods

In this phase, all the patent documents are classified concurrently by the following four classification methods: content-based patent classification, citation-based patent classification, metadata-based patent classification and patent network-based patent classification. The results generated by the four methods are then combined to yield the final patent classes as the output of this stage. In the following, we describe how the four methods are applied.

4.1.1 Content-Based Patent Classification

Previous studies reported that a patent's abstract is the most informative feature (Larkey, 1999; Liang, et al., 2003; Loh, et al., 2006). Thus we extract the content features from the titles and abstracts of the patent documents in this work. The details of content-based approach are described in section 2.1.1. After determining the similarity between the query patent and patents in the training patent dataset, the *k* nodes with highest similarity to the query patent document are extracted and used to identify the most appropriate class for a patent.

Under the content-based classification method, for a given query patent *q*, $F_{q,c}^{content}$ denotes the prediction score of query patent *q* belonging to class *c*. We choose the *k* nearest neighbor patents, S_q^{Nbr}, as references to calculate the prediction score, as shown in Eq. 11:

$$F_{q,c}^{content} = \sum_{p \in S_q^{Nbr}} \frac{B_{p,c}}{|S_q^{Nbr}|}, \tag{11}$$

$$\text{where } B_{p,c} = \begin{cases} 1, \text{if patent } p \text{ belongs to class } c \\ 0, \text{otherwise} \end{cases}$$

4.1.2 Citation-Based Patent Classification

Citation-based patent classification approaches include co-citation patent classification (Lai & Wu, 2005) and citation network patent classification (Li et al., 2007).

4.1.2.1 Co-citation patent classification

In the co-citation approach, the class of a query patent is determined by the majority vote of classes of its cited patents. The details of this approach are described in section 2.1.2.1.

For a given query patent *q*, let $F_{q,c}^{cocitation}$ denote the prediction score of query patent *q* belonging to class *c* under a citation-based classification method. The cited patents of *q*, S_q^{Cite}, are taken as references for calculating the prediction score, as shown in Eq. 12:

$$F_{q,c}^{cocitation} = \sum_{p \in S_q^{Cite}} \frac{B_{p,c}}{|S_q^{Cite}|}, \tag{12}$$

$$\text{where } B_{p,c} = \begin{cases} 1, \text{if patent } p \text{ belongs to class } c \\ 0, \text{otherwise} \end{cases}$$

4.1.2.2 Citation network patent classification

The details of the citation network approach (Li et al., 2007) are introduced in section 2.1.2.2.

By applying this approach, we retrieve two levels of cited patents from each patent document

to construct the citation network and train the classifier. The retrieved citation network of the set contains 25,348 patents in a citation network with 74 categories. Under the citation network classification method, for a given query patent q, $F_{q,c}^{citeNW}$ denotes the prediction score of query patent q belonging to class c, as defined in Eq. 13:

$$F_{q,c}^{citeNW} = SVM(q, sim_q, c) \ , \quad (13)$$

where sim_q denotes the vector of patent similarity between q and patents in the training set; sim_q=[$K(G_1,G_q)$, $K(G_2, G_q)$,…, $K(G_z,G_q)$], and z is the number of patents in the training set. Note that G_{pi} and G_{pj} represent the citation networks associated with two patents p_i and p_j; $K(G_{pi},G_{pj})$ denotes their patent similarity (Eq. 3). $SVM(q,sim_q,c)$ is the output of the SVM classifier for classifying q as of class c.

4.1.3 Metadata-Based Patent Classification

Richter and MacFarlane (2005) used metadata, such as inventors' names, to facilitate classification. More details of this approach are described in section 2.1.3. In this study, every patent document is represented by a vector of terms and inventors. After constructing the vectors, the similarity of two patent documents is calculated, and the *kNN* classifier is used to identify the appropriate class for the query patent based on the similarity (cosine value) of the patent documents.

Under the metadata-based classification method, for a given query patent q, $F_{q,c}^{metadata}$ denotes the prediction score of query patent q belonging to class c. We choose the k nearest neighbor patents, S_q^{Nbr}, as references to calculate the prediction score $F_{q,c}^{metadata}$ in Eq. 14.

$$F_{q,c}^{\text{metadata}} = \sum_{p \in S_q^{Nbr}} \frac{B_{p,c}}{|S_q^{Nbr}|} \quad (14)$$

.

4.1.4 Patent Network-Based Patent Classification

The proposed patent network-based approach constructs a patent ontology network based on the metadata of classified patents to represent the relationships among various field elements of the metadata. A query patent document can then be classified by searching for the "nearest" nodes in the patent ontology network, ranking them by their relevance scores and predicting the most appropriate class for the query patent. The approach involves four steps: patent ontology network construction, patent network analysis, k nearest neighbor extraction, and patent class identification. We described the steps in detail in Section 3.

The predicted score $F_{q,c}^{PNW}$ for a given query patent q belonging to class c is calculated using

Eq. 10, as illustrated in section 3.5.

4.2 Class combination

Under the proposed hybrid approach, each method generates a classification result based on the scores of the query patent in all candidate classes. The results generated by the four methods are then combined to yield the final patent classes as the output of this phase. Let $F_{q,c}^{\text{citation}}$ denote the prediction score of the citation-based patent classification, including co-citation approach (Eq. 12) and citation network approach (Eq. 13). The joint result $F_{q,c}$ is generated from the linear combination of $F_{q,c}^{\text{content}}$, $F_{q,c}^{\text{citation}}$, $F_{q,c}^{\text{metadata}}$ and $F_{q,c}^{PNW}$, as shown in Eq. 15:

$$F_{q,c} = \alpha \cdot F_{q,c}^{content} + \beta \cdot F_{q,c}^{citation} + \gamma \cdot F_{q,c}^{metadata} + \delta \cdot F_{q,c}^{PNW}, \quad (15)$$

where α, β, γ and δ are the respective weights of the four classification methods, which are determined empirically according to the best result in experiments.

The class with highest prediction score is then taken as the class of the query patent.

4.3 Experimental setup

4.3.1 Data collection

To evaluate the performance of the proposed approach, we conducted experiments on the collection of patent documents obtained from USPTO. The dataset contains 1,231 patent documents divided into 5 UPCs, as shown in Table 2. We use a patent's UPC to denote its class.

The documents in the database records are divided into two sets: 1) a training set (70% of the collected dataset) containing the patent documents whose classes are known; and 2) a test set (30% of the collected dataset) containing patent documents whose classes are to be determined.

Table 2. The collected patent dataset.

Class Number	Class Title	Data Instances
29	Metal Working	246
257	Active Solid-State Devices	273
324	Electricity: Measuring and Testing	221
438	Semiconductor Device Manufacturing Process	286
709	Electrical Computers and Digital Processing Systems: Multicomputer Data Transferring	205

4.3.2 Evaluation metrics

We used standard classification performance metrics, namely, the *accuracy rate*, *precision rate*, *recall rate*, and *F-measure* (Salton & Buckley, 1988; Van Rijsbergen, 1979), to evaluate the performance of the classifiers. These metrics have been widely used in information retrieval and machine learning studies.

Classification accuracy was used to assess the overall performance, as shown in Eq. 16:

$$Accuracy = \frac{\#\text{ of correctly classified patents}}{\text{total \# of patents}} \tag{16}$$

Precision, *recall* and *F-measure* were used to assess the classification performance. For instances of class *i*:

$$Precision\ (i) = \frac{\#\text{ of correctly identified patents for class } i}{\text{total \# of patents idientified as class } i} \tag{17}$$

$$Recall\ (i) = \frac{\#\text{ of correctly identified patents for class } i}{\text{total \# of patents in class } i} \tag{18}$$

Finally, to obtain a single performance measure, we used a simple *F-measure* to balance the *precision* and *recall scores*, as shown in Eq. 19:

$$F - measure\ (i) = \frac{2\times\text{precision } (i)\times\text{recall } (i)}{\text{presioion } (i)+\text{recall } (i)} \tag{19}$$

Precision and *recall* evaluate whether a classification is successful. If both parameters yield high scores in a classification experiment, the approach's performance is considered ideal. However, precision and recall are usually in conflict with each other, so the *F-measure* is used to balance the two results.

4.4 Experimental results and implications

4.4.1 Experiment one: link threshold of relevance calculation

The number of links in the patent network to expand has a significant effect on the results. The *k*-nearest neighbor extraction step attempts to identify the nodes that are most similar to the query patent document within the boundary defined by the given link threshold. If we limit expansion to only one link, all identified nodes have a direct relation to the query patent document. However, as the number of links increases, the number of nodes that have an

indirect link to the query patent will also increase.

Table 3 shows the performance of the patent network-based classification module under different link thresholds. The best performance is achieved when the link threshold = 3. Hence, we set the link threshold = 3 in the following experiments.

Table 3. The performance of the patent network-based classification module under different link thresholds

Link Threshold	Accuracy	Avg. precision	Avg. recall	Avg. F-measure
1	33.2	31.4	31.8	31.6
2	57.6	58.1	55.4	56.7
3	74.9	77.6	74.9	76.2
4	67.8	66.3	64.7	65.5

4.4.2 Experiment two: types of Nodes in the Patent Ontology Network (link threshold= 3)

The types of nodes used in the patent ontology network also affect the results. We tried to find the best types via experiments. As shown in Table 4, the patent ontology network with four types of nodes, namely, patent, class, inventor and assignee nodes, yields the best performance.

Table 4. The performance of the patent network with different combinations of nodes

Node types used	Accuracy	Avg. precision	Avg. recall	Avg. F-measure
Patent / class /inventor	61.9	68.8	65.3	67.0
Patent / class / assignee	68.5	66.1	71.4	68.6
Patent / class/ inventor / assignee	74.9	77.6	74.9	76.2

4.4.3 Experiment three: comparison of Different Patent Classification Methods

We compare four patent classification methods: content-based, citation-based, metadata-based and the proposed patent network-based classification methods. The content-based method described in Section 4.1.1 uses the similarity of content (title and abstract), and adopts the *kNN* classifier to predict the class of a query patent based on similarity measures of patents. The co-citation approach determines the class of a query patent by the majority vote of classes of its cited patents, as described in Section 4.1.2.1. The citation network approach described in Section 4.1.2.2 uses the similarity of citation network and employs an SVM classifier to predict the class of a query patent. We retrieve two levels of cited patents from each patent document to construct the citation network. The retrieved citation

network of the set contains 25,348 patents in a citation network. For metadata-based approach described in Section 4.1.3, the neighbors are chosen based on the similarities of the content (title and

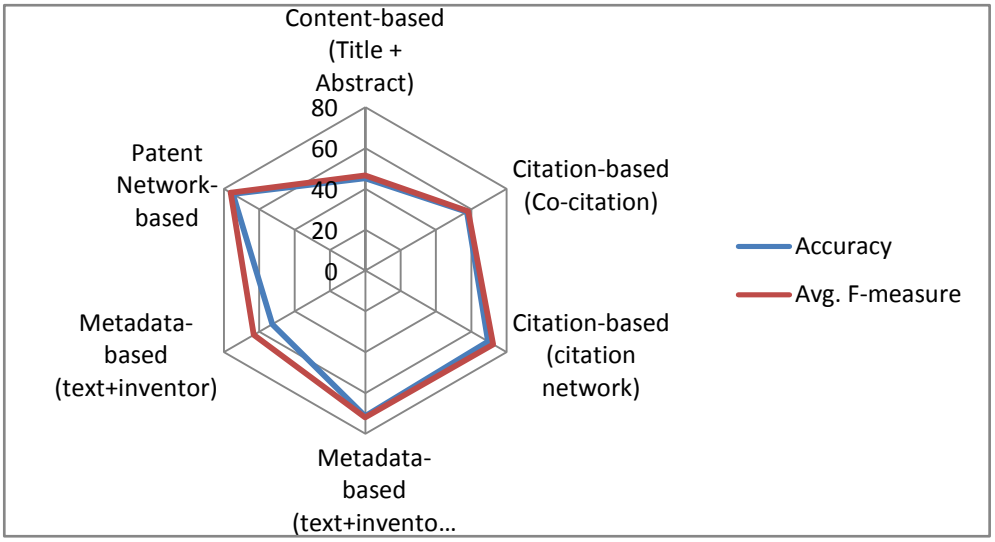

Figure 6. The performance of the compared patent classification methods

abstract), inventor and IPC. This approach also uses the *kNN* classifier to predict the class of a query patent. Note that our proposed patent network-based approach uses the relevance of nodes in the patent ontology network. A particular feature of the *kNN* classifier applied in our proposed patent network-based approach is that the neighbors can be of different types, such as patents and classes, whereas the other three methods only search for neighbors among patents.

Table 5 and Figure 6 show the performances of the compared patent classification approaches. The proposed patent network-based approach achieves the best performance in terms of accuracy (74.9%) and the F-score (76.2%). The second best approach, the metadata-based approach, considers the IPC when deciding the class of a query patent. The IPC denotes a kind of classification and may correlate with the UPC, which represents the class of a patent. Thus, it is not reasonable to consider IPC when making UPC class predictions. The Metadata-based (text + inventor + IPC) method may be affected by the correlation between IPC and UPC and thus yields a good result. Accordingly, we also compared the metadata-based approach without considering the IPC. The citation network approach performs better than the Metadata-based (text + inventor) method.

Table 5. The experiment results of the compared patent classification methods

Types of Patent Classification	Accuracy	Avg. precision	Avg. recall	Avg. F-measure
Content-based (Title + Abstract)	45.2	47.8	45.4	46.6
Citation-based (Co-citation)	57.6	54.2	62.8	58.2
Citation-based† (citation network)	69.5	71.4	73.5	72.4
Metadata-based (text+inventor+IPC)	71.3	75.6	68.7	72.0
Metadata-based* (text+inventor)	52.6	71.6	56.5	63.2
Patent Network-based	74.9	77.6	74.9	76.2

4.4.4 Experiment four: comparison of hybrid Patent Classification

In the proposed hybrid approach, each method generates a classification result and the joint result is derived by linear combination, as shown in Eq.15. Parameters α, β, γ and δ are the respective weights of the classification methods, which are determined empirically according to the best performance in experiments.

We choose the citation network method as the citation-based part of the proposed hybrid approach because it outperforms the co-citation method in the experiments, as mentioned in Sub-section 4.5.3. To avoid overlapping the effect of content-based part, we choose "Metadata-based (inventor)" as the metadata-based part of the proposed hybrid approach.

Table 6 shows the combination of different patent classification approaches and their weights. The goal of this experiment is to determine which combination of content, citation, metadata and patent network yields the best performance. The combination of the four methods (content-based, citation-based, metadata-based and patent network-based) achieves the best performance in terms of accuracy (84.1%) and the F-measure (86.4%). The weights of the four approaches are 0.1, 0.3, 0.1 and 0.5, respectively. In the experiments, we tested various combinations of α, β, γ and δ, by enumerating the values of the parameters in intervals of 0.1 ranging from 0 to 1, to find the best weight combination. For the hybrid effect, the result shows that the patent network-based method (with the highest weight 0.5) contributes the most in enhancing the performance of classification. The citation network method is more important than the content-based and metadata-based methods.

Table 7 shows the performances of the proposed hybrid approach and other patent classification methods The proposed hybrid approach with the weights α=0.1, β=0.3, γ=0.1 and δ = 0.5 achieves the best performance in terms of accuracy (84.1%) and the F-measure (86.4%). The second best approach is our proposed patent-network based method. The content-based method performs worse than other methods.

Table 6. The results of experiments using different combinations of patent classification approaches

Hybrid patent classification	α	β	γ	δ	Accuracy	Avg. F-measure
$\mathbf{C_{ontent}+C_{itation}+M_{etadata}^{*}+P_{atent\ network}}$	0.1	0.3	0.1	0.5	84.1	86.4
$\mathbf{C_{ontent}+C_{itation}+M_{etadata}^{*}}$	0.1	0.6	0.3	0	73.8	75.4
$\mathbf{C_{ontent}+C_{itation}+P_{atent\ network}}$	0.1	0.3	0	0.6	78.4	80.3
$\mathbf{C_{ontent}+M_{etadata}^{*}+P_{atent\ network}}$	0.1	0	0.2	0.7	77.0	78.8
$\mathbf{C_{itation}+M_{etadata}^{*}+P_{atent\ network}}$	0	0.3	0.2	0.5	83.2	86.2
$\mathbf{C_{ontent}+C_{itation}}$	0.2	0.8	0	0	71.9	74.2
$\mathbf{C_{ontent}+M_{etadata}^{*}}$	0.1	0.9	0	0	53.0	63.5
$\mathbf{C_{ontent}+P_{atent\ network}}$	0.1	0	0	0.9	75.5	78.5
$\mathbf{C_{itation}+M_{etadata}^{*}}$	0	0.7	0.3	0	73.5	75.2
$\mathbf{C_{itation}+P_{atent\ network}}$	0	0.4	0	0.6	76.4	79.4
$\mathbf{M_{etadata}^{*}+P_{atent\ network}}$	0	0	0.2	0.8	76.5	78.7

Table 7. The results of experiments using the hybrid approach and different patent classification methods

Types of Patent Classification	Accuracy	Avg. precision	Avg. recall	Avg. F-measure
Hybrid ($\mathbf{C_{ontent}+C_{itation}^{\dagger}+M_{etadata}^{*}+P_{atent\ network}}$)	84.1	85.2	87.7	86.4
Content-based (Title + Abstract)	45.2	47.8	45.4	46.6
Citation-based (Co-citation)	57.6	54.2	62.8	58.2
Citation-based† (citation network)	69.5	71.4	73.5	72.4
Metadata-based (text+inventor+IPC)	71.3	75.6	68.7	72.0
Metadata-based* (text+inventor)	52.6	71.6	56.5	63.2
Patent Network-based	74.9	77.6	74.9	76.2

Chapter 5 Patent Trends Change Mining

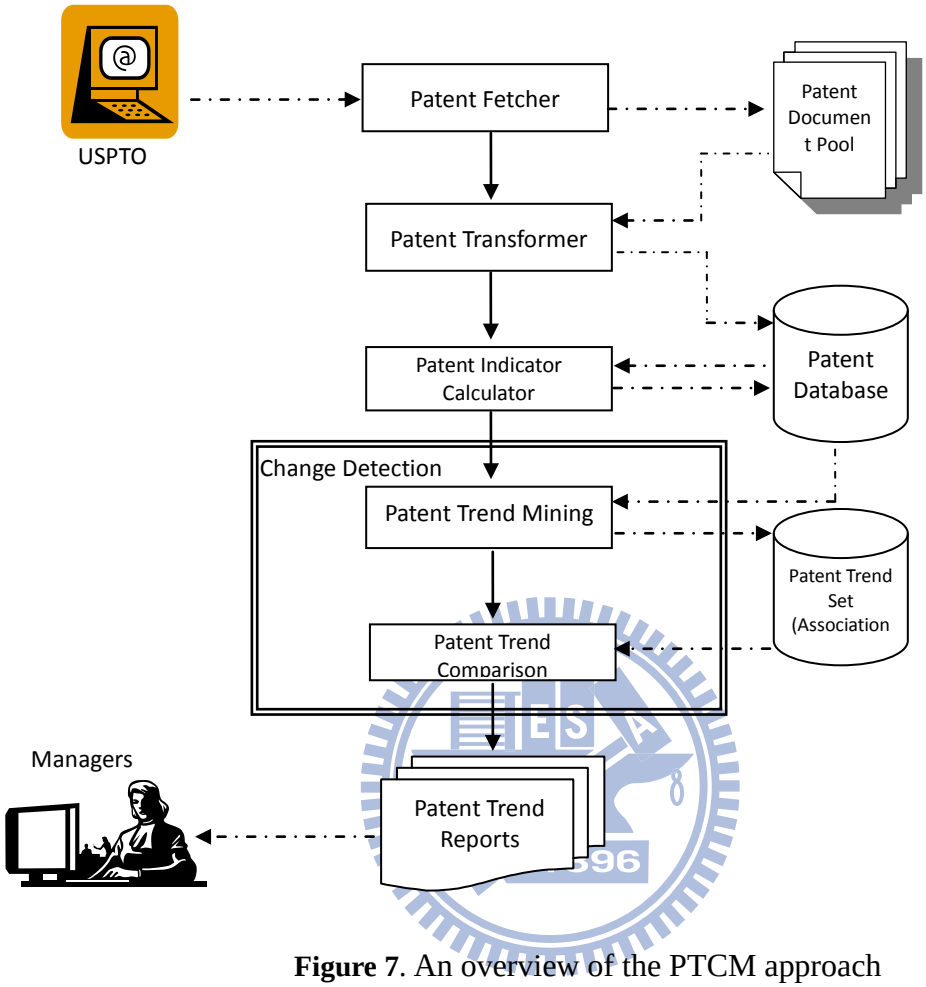

Figure 7. An overview of the PTCM approach

The proposed patent trend change mining (PTCM) approach (Shih et al., 2010) comprises four components, as shown in Figure 7: a patent fetcher, a patent transformer, a patent indicator calculator, and a change detection module. The first three components are described in this section, and we have more detail discussion on change detection process in Section 5.4.

5.1 Patent fetcher

With the rapid growth of computer and internet technologies, patent documents can now be accessed freely via the Internet. The patent fetcher module uses a keyword search strategy (e.g., Assignee and International Patent Classification Code (IPC)) to retrieve patents for analysis. Patent fetcher acquires patent documents (in HTML format) from the patent website and stores them into the patent document pool.

5.2 Patent transformer

Initially, a patent document is in a semi-structured HTML format. This module transforms the raw patent document from semi-structured HTML format into a text format, stores it in the database, filters out irrelevant content, and extracts required patent content, including the patent number, International Classification (IPC), Application Date, Assignee Name, and Assignee Country. The extracted content is stored in the database for further processing to compute patent indicators.

5.3 Patent indictor calculator

This module calculates the patent indicators for each patent to determine the patent's value. In this study we use four patent indicators, which are defined in Section 2.6, to analyze patent documents: *Citation Index* (CI) of a patent reflects the technological significance of a patent–the higher the value of a patent's CI, the greater the patent's impact. *Originality* measures the innovation of a patent–the higher the value of a patent's originality, the greater the patent's innovation value. *Generality* measures the scope of cross-field applications on which a patent is applied–the higher the value of a patent's generality, the greater the patent's economic value. A patent is interpreted as having more "generality" if the forward citations are spread over several technological fields. *Technology Cycle Time* (TCT) measures the time between the previous patent and the target patent, which makes improvement on the previous one–shorter TCT means a faster technological progress of patents.

The values of patent indicators are discretized for further patent trend mining. We perform data discretization based on the normalized results derived by SPSS Visual Bander. The values of patent indicators are transformed into linguistic terms as shown in Table 8.

Table 8. Data discretization of patent indicators

Patent Indictor	Linguistic term	Numerical range
CI	Low	≤0
	Mid	1-4
	High	≥5
Originality	Low	0-0.39
	Mid	0.40-0.65
	High	0.66-1

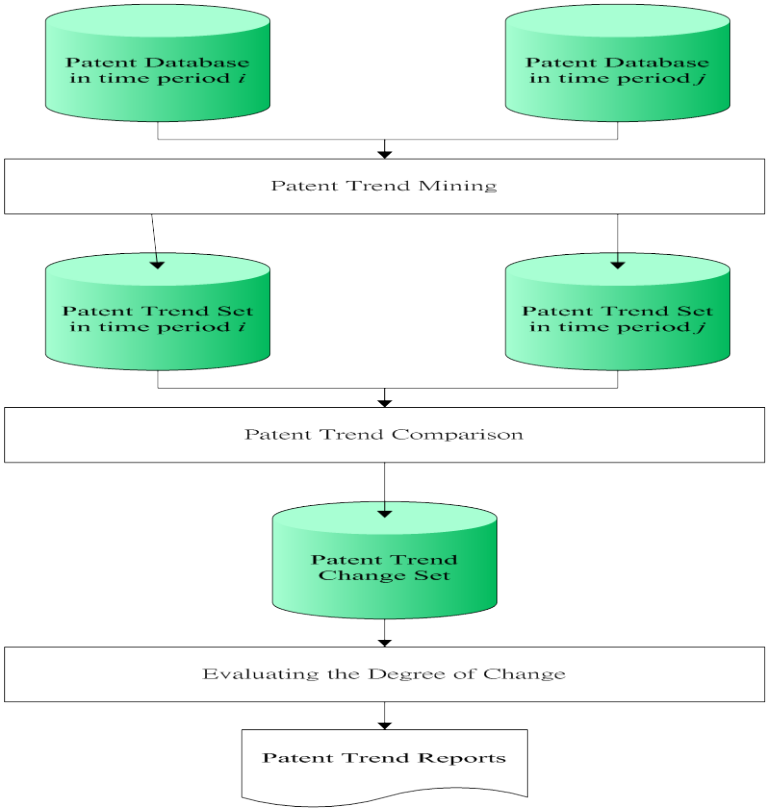

Figure 8. The process of detecting changes in patent trends

Generality	Low	0-0.44
	Mid	0.45-0.65
	High	0.66-1
TCT	Short	0-5
	Mid	6-7
	Long	≥8

5.4 Change detection in patent trends

Patents indicate the technological competitiveness as well as the innovation strategy of a company in a given period. Business managers can observe changes in patent trends by comparing the trends of two periods. The process of detecting changes in patent trends is illustrated in Figure 8.

5.4.1 Patent trend mining

Before describing the patent trend mining module, we introduce the patent trends analyzed in this study. To assist business executives in understanding trends in the development of

technologies and plan suitable strategies, we define four kinds of patent trends and classify them into two levels for analysis: company-level and industry-level trends.

(a) company-level patent trends: these trends provide information about a company's technological development.

- Trends in the R&D activities of a company: changes in the R&D activities of a company can be determined by comparing the relations between technological fields (IPC) and four patent indicators (the citation index, originality, generality and technology cycle time described in Section 5.3) over two time periods.

(b) industry-level patent trends: these trends provide information about the technological development of an industry.

- Trends in the R&D activities of an industry: changes in the R&D activities of an industry can be determined by comparing the relations between the technological fields (IPC) and four patent indicators over two time periods.

- Trends in the technological competitiveness of companies: we identify changes in technology competitiveness of companies by comparing the relations between a patent's assignee (company) and the four patent indicators over two time periods; the patent indicators reflect the technological competitiveness of a company.

- Trends in the technological competitiveness of companies in a specific technological field: these changes can be observed by comparing the relations between both a patent's assignee and technological fields (IPC) and four patent indicators over two time periods.

Table 9 shows the four kinds of patent trends and their respective rule formats.

Table 9. Patent trends and their respective rule formats

Analyzed level	Patent trend	Rule format		
		Conditional part	→	Consequent part
Company level	R&D activities of a company	IPC	→	CI/ Originality/ Generality/ TCT
Industry level	R&D activities of the specified industry	IPC	→	CI/ Originality/ Generality/ TCT
	Technological competitiveness of companies	Assignee	→	CI/ Originality/ Generality/ TCT
	Technological	Assignee, IPC	→	CI/ Originality/

	competitiveness of companies in a specific technological field			Generality/ TCT

We apply association rule mining to patent data to identify patent trends (frequent association rule patterns). The mined frequent patterns can be regarded as trends extracted from patent documents. For example, if there are sufficient patents belonging to technological field *B*, whose assignee is *X*, and the CI value of those patents is high, the frequent association rule pattern "Assignee= *X*, IPC=*B*→ CI= high" can be identified. The rules identify a patent trend in which the citation index of X's patents in technological field *B* is relatively high. This information suggests that the quality of *X*'s patents in technological field B is high in the industry. Moreover, we may say that *X* is a pioneer company in technological field *B*.

5.4.2 Patent trend comparison

After the patent trends of different time periods have been discovered, the trends (in rule format) are compared to identify changes. We start with defining the types of change as follows and then discuss the process of trend comparison.

5.4.2.1. Types of change

Based on previous research (Song, Kim & Kim, 2001), four types of change in patent trends are defined:

(1) Emerging patent trends: an emerging patent trend is a rule pattern whose support increases significantly from one dataset to another.

(2) Unexpected changes in patent trends: unexpected changes in patent trends can be found in newly discovered patent trends whose consequent parts of the rule patterns are different from those of the previous patent trend.

(3) Added patent trends: an added patent trend is a new rule, i.e., a rule not found in previous rule patterns.

(4) Perished patent trends: a perished patent trend is the opposite of an added rule, as it is only found in previous rule patents.

5.4.2.2. Rule matching

We use a rule matching method to compare the patent trends of different time periods. The method computes the similarity measures and difference measures of the patent trends $rule_i^t$ and $rule_j^{t+k}$ in time t and time $t+k$, respectively. The modified rule matching method comprises the following four steps (Liu et al. 2009; Song, Kim & Kim, 2001).

Step 1. Calculate the similarity degree of the conditional / consequent parts of two rules in different time periods.

Step 2. Calculate the similarity measure S_{ij} between two rules. The measure is derived by multiplying the similarity degree of the conditional parts (C_{ij}) of the rules by the similarity degree of the consequent parts (Q_{ij}).

Step 3. Calculate the difference measure ∂_{ij} between two rules. The measure is the similarity degree of the conditional parts minus the similarity degree of the consequent parts.

Step 4. Determine the type of change according to the similarity measures and difference measures.

5.4.2.3. Identifying the type of change

Table 10 shows the measures used to determine each type of event change; the measurements are adopted from (Liu et al. 2009; Song, Kim & Kim, 2001). The four types of event change can be classified according to the two judged factors, i.e., the similarity measure S_{ij} and the difference measure ∂_{ij}, and three predefined thresholds: θ_{em} for emerging patterns, θ_{un} for unexpected changes, and $\theta_{a/p}$ for added and perished rules. Note that $\theta_{em} > \theta_{un} > \theta_{a/p}$. The process of identifying the types of changes follows a pre-determined sequence. First, we identify emerging patterns. If the similarity measure S_{ij} is greater than or equal to θ_{em}, it means that the two rules are similar and rule r_j^{t+k} can be regarded as an emerging pattern. If the maximum similarity measure $Max(\varsigma_i, \varsigma_j)$ is less than θ_{em} and the difference measure ∂_{ij} is greater than θ_{un}, we regard rule r_j^{t+k} as an unexpected change. Note that $\varsigma_i = \max_j S_{ij}$; $\varsigma_j = \max_i S_{ij}$. Finally, if ς_j is less than $\theta_{a/p}$, rule r_j^{t+k} is identified as an added patent trend; and if ς_i is less than $\theta_{a/p}$, rule r_i^t is identified as a perished patent trend.

5.4.3 Evaluating the degree of change

As a large number of changes occur in a competitive business environment, managers need to focus on the most important changes. To do this, it is necessary to evaluate the degree of change, and rank the changed rules according to their importance.

Table 10. Measurement for each type of change

Type of Change (r_i^t , r_j^{t+k})	Measurement
Emerging Pattern	$S_{ij} \geq \theta_{em}$ ($S_{ij} = C_{ij} \times Q_{ij}$) (C_{ij} : similarity degree of the conditional parts) (Q_{ij} : similarity degree of the consequent parts)
Unexpected Change	$Max(\varsigma_i, \varsigma_j) < \theta_{em}$, $\partial_{ij} > \theta_{un}$ ($\partial_{ij} = C_{ij} - Q_{ij}$)
Added Patent trend	$\varsigma_j < \theta_{a/p}$ ($\varsigma_j = \max_i S_{ij}$)
Perished Patent trend	$\varsigma_i < \theta_{a/p}$ ($\varsigma_i = \max_j S_{ij}$)

Table 11. Measuring the degree of change in patent trends

Type of Change	Degree of Change
Emerging patent trends	$\frac{Support^{t+k}(r_j) - Support^t(r_i)}{Support^t(r_i)}$
Unexpected changes in patent trends	$\frac{Support^t(r_i) - Support^{t+k}(r_i)}{Support^t(r_i)} \times Support^{t+k}(r_j)$
Added patent trend	$(1-\varsigma_j) \times Support^{t+k}(r_j)$
Perished patent trend	$(1-\varsigma_i) \times Support^t(r_i)$

Table 11 shows the simple formulations for measuring the degree of change. The formulations, which are adopted from (Liu et al. 2009), measure the degree of change. The notations support^t(r_i) and support^{t+k}(r_i) represent the support value of r_i at time t and and r_j at time $t+k$, respectively; while ς_i and ς_j are the maximum similarity measures of r_i^t and r_j^{t+k}, respectively.

After calculating the degrees of change, the most important changes are reported to business managers, who then analyze the changes in patent trends over different time periods and use the information to understand the changing business environment and plan appropriate strategies.

5.5. Experimental setup

5.5.1 Data collection

The dataset of semiconductor-related patents was obtained from the USPTO (United States Patent and Trademark Office) patent database. We select Taiwan semiconductor-related patents available online for the period 2001-2004 based on the IPCs belonging to the semiconductor industry, as identified by the Taiwan Intellectual Property Office (see Appendix B). We divided this dataset, which contains 4,310 unique patents, into two periods: the first part contains 2,352 patent documents for the period 2001 to 2002, while the second part contains 1,958 patent documents for the period 2003 to 2004.

5.6. Experimental results and implications

5.6.1 Experiment one: Changes in the R&D activities of TSMC (Taiwan Semiconductor Manufacturing Co. Ltd)

Changes in a company's R&D activities are identified by comparing the relations between the technological field (IPC) of the target company and the citation index, originality, generality, and technology cycle time over two time periods. We chose TSMC as the target company, and divided its patents into two parts: 2001-2002 and 2003-2004. Table 12 lists some changes in the R&D activities of TSMC between 2001 and 2004.

Table 12. Some changes in the R&D activities of TSMC

Patent trend		Change Degree
Emerging patent trends		
(1) IPC=H01L29/788 → Originality= High		0.57
(2) IPC=H01L21/00 → TCT= Short		0.21
Unexpected changes in patent trends		
2001-2002	**2003-2004**	
(3) IPC=H01L27/108→ CI= Mid	IPC=H01L27/108 → CI=Low	0.02
(4)) IPC=H01L21/311→ TCT= Short	IPC=H01L21/311→ TCT= Long	0.02
Added patent trends		
(5) IPC=H01L23/62 → CI= Low		0.03
(6) IPC=G01R31/26 → CI=Low		0.02
Perished patent trends		
(7) IPC=H01L21/336 → CI= High		0.05
(8) IPC=H01L21/44 → Generality=High		0.03

From patent trend (1), we observe the rapid growth (57%) of the company in terms of high originality in H01L29/788. This information shows that, during the period under study, TSMC exhibited a high degree of inventiveness in the technological field H01L29/788.

Meanwhile, patent trend (3) shows that the citation index of H01L27/108 decreased between 2001 and 2004. A reduction in the CI often indicates a decline in quality, although it can mean that the patent is fairly new. The added patent trends (5) and (6) in Table 12 indicate that H01L21/336 and G01R31/26 are new technological fields that TSMC invested in. The number of citations of these patents is relatively low. Finally, from perished patent trends (7) and (8), we observe that the innovativeness of TSMC declined gradually in terms of H01L21/336 and H01L21/44 in the period under study.

5.6.2 Experiment two: changes in the R&D activities of Taiwan's semiconductor industry

Changes in the R&D activities of an industry are identified by comparing the relations between the technological fields (IPC) of the target industry and the citation index, originality, generality and technology cycle time over two time periods. Table 6 lists some changes in the R&D activities of Taiwan's semiconductor industry between 2001 and 2004.

In Table 13, the emerging patent trends (1) and (2) show that companies in the industry invested in H01L29/76 and H01L21/00 consistently throughout the period under study. The high growth rates (131% and 107% respectively) indicate that companies focused their R&D activities on the two technological fields. However, the low CI indicates that the companies lacked pioneer patents and basic patents in these technological fields.

Patent trends (3) and (4) in Table 13 indicate that the TCT of H01L29/40 and H01L21/48 changed from a short-cycle time to a medium-cycle time, which implies that the speed of innovation in these technological fields slowed down. The added patent trends (5) and (6) indicate that H01L29/788 and G11C16/04 were new technological fields that Taiwanese semiconductor companies invested in during 2003-2004.

Table 13. Some changes in the R&D activities of Taiwan's Semiconductor Industry

Patent trend		Change Degree
Emerging patent trends		
(1) IPC=H01L29/76 → CI= Low		1.31
(2) IPC=H01L21/00 → CI= Low		1.07
Unexpected changes of patent trends		
2001-2002	**2003-2004**	

(3) IPC=H01L29/40 → TCT= Short	IPC=H01L29/40 → TCT= Mid	0.02
(4) IPC=H01L21/48→ TCT= Short	IPC=H01L21/48→ TCT= Mid	0.01
Added patent trends		
(5) IPC=H01L29/788 → CI= Low		0.03
(6) IPC=G11C16/04 → CI= Low		0.02

5.6.3 Experiment three: Technological competitiveness of companies in Taiwan's semiconductor industry

Changes in the technological competitiveness of companies in an industry are identified by comparing the relations between the assignee of the target industry and the citation index, originality, generality, and technology cycle time over two time periods. Table 14 lists some changes in the technological competitiveness of companies in Taiwan's semiconductor industry between 2001 and 2004.

Patent trends (1) and (2) in Table 14 show the consistent innovative power of TSMC and MIC. Specifically, the marked increase in MIC's patents (263%) indicates the innovativeness of MIC and the direction of its R&D activities. However, the low CI indicates that MIC was a technological follower between 2001 and 2004. Patent trend (4) in Table 8 shows a decrease in the Originality of SPIC. The added patent trends (5) and (6) in the table show several new assignees of semiconductor patents, which means that new companies (AOC and NYT) entered the semiconductor industry during 2003-2004.

From the perished patent trends (7) and (8), we observe that the high value of CI and the Generality of TSMC's patents decreased between 2003 and 2004. This implies that the quality of TSMC's R&D may have declined during 2003-2004, although the phenomenon may be due to new patents.

Table 14. Some changes in the technological competitiveness of companies in Taiwan's semiconductor industry

Patent trend		Change Degree
Emerging patent trends		
(1) Assignee=Macronix International Co. Ltd→ CI= Low		2.63
(2) Assignee= Taiwan Semiconductor Manufacturing Co. Ltd → Originality= High		0.01
Unexpected changes in patent trends		
2001-2002	**2003-2004**	
(3) Assignee= Advanced Semiconductor Engineering, Inc. → CI= High	Assignee= Advanced Semiconductor Engineering, Inc. → CI= Low	0.32

(4)) Assignee=Siliconware Precision Industries Co., Ltd.→ Originality= High	Assignee=Siliconware Precision Industries Co., Ltd.→ Originality= Low	0.03
Added patent trends		
(5) Assignee=Au Optronics Corp. → CI= Low		0.04
(6) Assignee=Nan Ya Technology → CI=Low		0.03
Perished patent trends		
(7) Assignee= Taiwan Semiconductor Manufacturing Co. Ltd → CI= High		0.07
(8) Assignee= Taiwan Semiconductor Manufacturing Co. Ltd → Generality= Mid		0.07

5.6.4 Experiment four: Technological competitiveness of companies in specific technological fields

Changes in the technological competitiveness of companies in specific technological fields are derived by comparing the relations between both the patent's assignee and the technological field (IPC) of the target industry with the citation index, originality, generality, and technology cycle time over two time periods. Table 15 lists some changes in Taiwan's semiconductor industry between 2001 and 2004.

The frequent appearance of TSMC in emerging patent trends shows that the company played a leading role in Taiwan's semiconductor industry throughout the period under study. The perished patent trends (3) and (4) in Table 15 show that UMC's technological competitiveness with medium CI and low Originality in H01L21/336 declined, which may imply a change in UMC's innovative activities.

Table 15. Some changes in the activities of Taiwan's semiconductor industry

Patent trend	Change Degree
Emerging patent trends	
(1) IPC=H01L21/302, Assignee=Taiwan Semiconductor Manufacturing Co. Ltd → CI= Low	1.4
(2) IPC=H01L21/44, Assignee=Taiwan Semiconductor Manufacturing Co. Ltd → CI= Low	0.78
Perished patent trends	
(3) IPC=H01L21/336, Assignee= United Microelectronics Corp. → CI= Mid	0.02
(4) IPC=H01L21/336, Assignee= United Microelectronics Corp. → Originality= Low	0.02

Chapter 6 Concluding Remarks

In this thesis, we proposes two approaches for different patent analysis purpose: the hybrid patent classification approach for automatically categorizing patents, and the patent trend change mining approach for detect technological change trends.

We have proposed a novel patent network-based classification method, which uses patent metadata to derive the weights of the relationships between different types of nodes in the patent network, Based on the patent network analysis, the classification result can be improved by considering the neighboring patent nodes and class nodes of a query patent in making class prediction. The main contributions of the proposed method include novel designs on (a) patent network construction based on the proposed relationship metrics between different types of patent nodes; and (b) patent class prediction based on the patent network analysis and the modified *kNN* classifier. Our experiment results demonstrate that the proposed patent network-based method outperforms the content-based, citation-based and metadata-based methods. Moreover, we combine the patent network-based method with three conventional classification methods to develop a hybrid patent classification approach. Our experiment results demonstrate that the hybrid approach performs better than the patent network-based method. The proposed hybrid patent classification approach can further enhance the classification performance by a hybrid of multiple classifiers. For the hybrid effect, the result shows that the patent network-based method is more important than other methods in enhancing the performance of classification.

The proposed patent trend change mining approach captures changes in patent trends without the need for specialist knowledge and reports changes to business managers by ranking the degrees of change. Competitive intelligence of business is derived by an automatic change mining approach that business managers can modify and develop appropriate strategies according to their findings. The proposed approach mines changes in patent trends by analyzing the metadata in patent documents. We applied the proposed PTCM to Taiwan's semiconductor industry for the period 2001-2004 to discover changes in four types of patent trends: the R&D activities of a company, the R&D activities of the industry, the technological competitiveness of companies and the technological competitiveness of companies in a specific technological field. The results obtained by the proposed approach can be used as an important reference for decision makers to make more accurate strategies on research and development.

There remain several extended researches to do base on this study. The primary part of most patent document is textual content which contains rich information to utilize (e.g., abstracts and claims). Through analyzing the textual part we can surely improve the quality of change detection and provide more comprehensive results. Therefore the next research will be a patent trend change mining approach which utilizes text mining techniques.

An obvious task to put effort on is to find out the best combination of weights for hybrid patent classification, i.e, parameter α, β, γ and δ in Eq. 15. The combination might change depending on target industries, data fields to analysis and terminology distribution, etc. We have designed a series of experiments to find out the best weight combination for the example in this thesis. And in fact, how to accurately determine the best weight combination for various cases will also be an issue to study.

Another important future work is to develop an effective validation approach for examining the results obtained from patent trend change mining, and for conducting further analysis. Traditional indices for evaluating data mining results might not be adequate for patent trend change mining. Approaches which evaluate the results more accurately can significantly improve the precision and accuracy of the model. This is also a challenging work for us to do.

References

Agrawal, R., Srikant, R. (1994). Fast algorithm for mining association rules. *Proc Int Conf on Very Large Data Bases*, Santiago de Chile. San Francisco, CA: Morgan Kaufmann, 487-499.

Berio, G., Harzallah, M. (2007). Towards an integrating architecture for competence management. Computers in Industry, 58 (2), 199-209.

Breitzman, A. F., Mogee, M. E. (2002). The many applications of patent analysis. Journal of Information Science,28 (3), 187-205.

Brockhoff, K. K. (1991). Indicators of firm patent activities. In: Technology Management: the New International language, 476-481.

Chang, M. C. (2005). Quantum computation patent mapping- a strategic view for the information technique of tomorrow. International Conference on Service Systems and Service Management, 1177-1181.

Chang, C.-W., Lin, C.-T., and Wang L.-O. (2009). Mining the text information to optimizing the customer relationship management. Expert Systems with Applications, 36 (2), 1433-1443.

Chen, M. C., Chiu, A. L., Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. Expert Systems with Applications, 28 (4), 773-781.

Chen, S. Y., Liu, X. (2004). The contribution of data mining to information science. Journal of Information Science, 30 (6), 550-558.

CHI-Research, http://www.chiresearch.com.

Cong, H. and Tong, L. H. (2008). Grouping of TRIZ Inventive Principles to facilitate automatic patent classification. Expert Systems with Applications, 34, 788-795.

Cong, H. and Loh, H. T. (2010). Pattern-Oriented Associative Rule-Based Patent Classification. Expert Systems with Applications, 37(3), 2395-2404.

Dou, H., Leveillé, V., Manullang, S. & Dou, J. J. (2005). Patent analysis for competitive technical intelligence and innovative thinking. Data Science Journal, 4, 209-237.

Dürsteler, J. C. Patent analysis. http://www.infovis.net/ , visited 2007/5.

EPO Web Site, http://www.european-patent-office.org/index.en.php , European Patent Office.

Fall, C. J., Torcsvari, A., Benzineb, K. and Karetka, G. (2003). Automated categorization in the International Patent Classification. In SIGIR Forum, 10-25.

Fall, C. J., Torcsvari, A., Benzineb, K. and Karetka, G. (2004). Automated categorization of German-language Patent Documents. Expert Systems with Applications, 26(2), 269-277.

Guan, J. C. and Gao, X. (2009). Exploring the h-Index at Patent Level. Journal of the American Society for Information Science and Technology, 60(1), 35-40.

Han, J., Kamber, M. (2001). Mining association rules in large databases. Data Mining-Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.

Huang, Z., Chen, H., Yip, A., Ng, G., Guo, F., Chen, Z. K., Roco, M. C. (2003). Longitudinal patent analysis for nanoscale science and engineering: country, institution and technology field. Journal of Nanoparticle Research, 5 (3-4), 333-363.

Huang, S.-H., Ke, H.-R., Yang, W.-P. (2008). Structure clustering for Chinese patent documents. Expert Systems with Applications, 34 (4), 2290-2297.

Ian, H. W., Eibe, F. (2000). Output: Knowledge Representation. Data Mining. San Francisco: Morgan Kaufmann Publishers.

Kim, J. H. and Choi, K. S. (2007). Patent document categorization based on semantic structural information. Information Processing and management, 43, 1200-1215.

Kim, Y. G., Suh, J. H., Park, S. C. (2008). Visualization of patent analysis for emerging technology. Expert Systems with Applications, 34 (3), 1804-1812.

Kohonen, T., Kaski, S., Lagus, K., Salojavi, J., Honkela, J., Paatetro, V., et al. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11(3), 574-585.

Kuo, R. J., Lin, S. Y., Shih, C.W. (2007). Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan. Expert Systems with Applications, 33 (3), 794-808.

Lai, K. K., Wu, S. J. (2005). Using the Patent Co-citation Approach to Establish a New Patent Classification System. Information Processing and Management, 41, 313-330.

Larkey, L. S. (1999). A Patent Search and classification system. In Proceedings of the fourth ACM conference on Digital libraries, 179-183.

Li, X., Chen, H. C., Zhang, Z., Li, J. (2007). Automatic Patent Classification using Citation Network Information: An Experimental Study in Nanotechnology. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, 419-427.

Liu, B., Hsu, W., Ma, Y. (2001). Discovering the set of fundamental rule changes. Proc. of the 7th ACM International Conference on Knowledge Discovery and Data Mining. San Francisco, California, 335-340.

Liu, B., Hsu, W. (1996). Post-analysis of learned rules. Proc. of 13th National Conference on Artificial Intelligence, Menlo Park, California, 828-834.

Liu, B., Hsu, W., Han, H. S., Xia, Y. (2000). Mining changes for real-life applications. Proc. of the 2nd Int Conf on Data Warehousing and Know Discovery, London, 337-346.

Liu, D. R., Shih M. J., Liau C. J. & Lai, C. H. (2009). Mining the change of event trends for decision support in environmental scanning. Expert Systems with Applications, 36 (2), 972-984.

Loh, H. T., He, C. and Shen, L. (2006). Automatic classification of patent documents for TRIZ users. World Patent Information, 28(1), 6-13.

Ngai, E.W.T., Xiu L., and Chau, D.C.K., (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems

with Applications, 36 (2), 2592-2602.

O'Hara, K. Alani, H. and Shadbolt, N. (2002). Identifying Communities of Practice: Analysing Ontologies as Network to Support Community Recognition. In Proceeding Conference International Federation Information Processing, World Computer Congress.

Reitzig, M. (2004). Improving patent valuations for management purposes- validating new indicators by analyzing application rationales. Research Policy , 33 (6-7), 939-957.

Richter, G. and MacFarlane, A. (2005). The Impact of metadata on the accuracy of automated patent classification. World patent Information, 27(1), 13-26.

Salton, G. and Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. Information Process Management, 24(4), 323-328.

Shih, M. J. and Liu, D. R. (2010). Patent classification using ontology-based patent network analysis. In Pacific Asia Conference on Information System (PACIS 2010), July 9-12, 2010, Taipei, Taiwan.

Shih, M. J., Liu, D. R., and Hsu, M. L. (2010). Discovering competitive intelligence by mining changes in patent trends. Expert Systems with Applications, 37, 2882-2890.

Song, H. S., Kim, J. K., Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. Expert Systems with Applications, 21 (3), 157-168.

Stembridge, B. (2005). Sorting the wheat from the chaff- the use of patent analysis in evaluating portfolios. http://www.scientific.thomson.com/newsletter.

Stembridge, B., Corish, B. (2004). Patent data mining and effective patent portfolio management. Intellectual Asset Management, Oct./Nov, 30-35.

Su, F. P., Lai, K. K., Sharma, R. R. K. and Kuo, T. H. (2009). Patent priority Network: Linking Patent Portfolio to Strategic Goals. Journal of the American Society for Information Science and Technology, 60(11), 2353-2361.

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

Tuomo, N., Hermans, R., Kulvik, M. Patent citations indicating present value of the biotechnology business. http://www.etla.fi/.

USPTO Web Site, http://www.uspto.gov/ , United States Patent & Trademark Office.

Van Rijsbergen, C. J. (1979). Information retrieval (2nd ed.). London: Butterworths.

WIPO Web Site, http://www.wipo.int/portal/index.html.en , World Intellectual Property Organization.

Yang, Y. (1994). Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorisation and Retrieval. In: Croft WB, van Rijsbergen CJ, editors. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 3-6 July. New York: ACM, 13-22.

Yang, Y. Y., Akers, L., Klose, T., Yang, C. B. (2008). Text mining and visualization tools –

Impressions of emerging capabilities. World Patent Information, 30 (4), 280-293.

Yen, S. J., Lee, Y. S. (2006). An efficient data mining approach for discovering interesting knowledge from customer transactions. Expert Systems with Applications, 30 (4), 650-657.

Yuan, Y. C., Carboni, I. and Ehrlich K. (2010). The Impact of Awareness and Accessibility on Expertise Retrieval: A Multilevel Network Perspective. Journal of the American Society for Information Science and Technology, 61(4), 700-714.

Appendix A.

The algorithm of patent network analysis.

```
Initialize all nodes weights to 1
Create a relationship-array of relationships and weights
Set query patent document as the active node
Mark current node as unlocked and add it to a node-array
Loop to the maximum number of links to traverse
    Search for the current node in node-array
    If found:
        Mark node as locked
        Set node as the active node
        Get all node connected to current node with a relationship in the
        relationship-array
        Loop to number of connected nodes
            If node not in node-array (new node)
                Weight of node=initial weight + current node weight
                                * weight of connecting relation
            Mark node as unlocked and add it to node-array
            If node already in node-array
                Weight of node=node weight + current node weight
                                * weight of connecting relation
        End loop
    If not found then exit
End loop
Relevance of node = Weight of node / n
(n= the minimum number of the links traversed to reach the node starting from the
query node)
```

Appendix B.

IPCs belonging to the semiconductor industry identified by the Taiwan Intellectual Property Office

IPC	Description
C23C	Coating metallic material; Coating material with metallic material; Surface treatment of metallic material by diffusion into the surface, by chemical conversion or substitution; Coating by vacuum evaporation, by sputtering, by ion implantation or by chemical vapor deposition
016/00	Chemical coating by decomposition of gaseous compounds, without leaving reaction products of surface material in the coating, i.e. chemical vapor deposition (CVD) processes
G01R	Measuring electric variables; Measuring magnetic variables
031/02	General constructional details
G03F	Photomechanical production of textured or patterned surfaces, e.g., for printing, for processing of semiconductor devices;
007/00	Photomechanical, e.g., photolithographic, production of textured or patterned surfaces, e.g., printed surfaces;
009/00	Registration or positioning of originals, masks, frames, photographic sheets, or textured or patterned surfaces
G05F	Systems for regulating electric or magnetic variables
001/10	Regulating voltage or current
G11C	Static stores
007/00	Arrangements for writing information into, or reading information from, a digital store
016/04	Using variable threshold transistors, e.g., FAMOS
H01L	Semiconductor devices; Electronic solid state devices
021/00	Processes or apparatus specially adapted for the manufacture or treatment of semiconductor or solid state devices or parts thereof
023/34	Arrangements for cooling, heating, ventilating or temperature compensation
023/48	Arrangements for conducting electric current to or from the solid state body in operation, e.g., leads, terminal arrangements
023/495	Lead-frames
023/52	Arrangements for conducting electric current within the device in operation from one component to another
023/58	Structural electrical arrangements for semiconductor devices
023/62	Protection against over-current or overload, e.g., fuses
027/108	Dynamic random access memory structures
029/00	Semiconductor devices specially adapted for rectifying, amplifying, oscillating or switching and having at least one potential-jump barrier or surface barrier**;** Capacitors or resistors with at least one potential-jump barrier or surface barrier, e.g. PN-junction depletion layer or carrier concentration layer**;** details of semiconductor bodies
029/40	Electrodes
029/76	Unipolar devices
029/788	With floating gate
029/94	Metal-insulator-semiconductors, e.g., MOS
031/062	The potential barriers being only of the metal-insulator-semiconductor type

031/113	Being of the conductor-insulator- semiconductor type, e.g., metal-insulator-semiconductor field-effect transistor
031/119	Characterized by field-effect operation, e.g., MIS type detectors

Appendix C.

Normative sections of patent documents.