Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Information Processing and Management 41 (2005) 313–330
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Using the patent co-citation approach to establish a new patent classification system

Kuei-Kuei Lai [a,*], Shiao-Jun Wu [b,c]

[a] *Department of Business Administration, National Yunlin University of Science and Technology, Touliu, Yunlin 640, Taiwan, ROC*
[b] *Department of Information Management, Kun Shan University of Technology, Taiwan, ROC*
[c] *Graduate School of Management, National Yunlin University of Science and Technology, Taiwan, ROC*

Received 25 August 2003; accepted 19 November 2003
Available online 23 December 2003

Abstract

The paper proposes a new approach to create a patent classification system to replace the IPC or UPC system for conducting patent analysis and management. The new approach is based on co-citation analysis of bibliometrics. The traditional approach for management of patents, which is based on either the IPC or UPC, is too general to meet the needs of specific industries. In addition, some patents are placed in incorrect categories, making it difficult for enterprises to carry out R&D planning, technology positioning, patent strategy-making and technology forecasting. Therefore, it is essential to develop a patent classification system that is adaptive to the characteristics of a specific industry. The analysis of this approach is divided into three phases. Phase I selects appropriate databases to conduct patent searches according to the subject and objective of this study and then select basic patents. Phase II uses the co-cited frequency of the basic patent pairs to assess their similarity. Phase III uses factor analysis to establish a classification system and assess the efficiency of the proposed approach. The main contribution of this approach is to develop a patent classification system based on patent similarities to assist patent manager in understanding the basic patents for a specific industry, the relationships among categories of technologies and the evolution of a technology category.

Keywords: Patent management; Patent classification system; Co-citation; Bibliometrics

* Corresponding author. Tel.: +886-5-53-42-601x5228; fax: +886-5-53-12-074.
E-mail address: laikk@yuntech.edu.tw (K.-K. Lai).

0306-4573/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2003.11.004

1. Introduction

A patent is a contract between an inventor and the government, whereby in return for full public disclosure of an invention, the government grants the inventor the right to exclude others for a limited time from making, using and selling the invention (Hufker & Alpert, 1994). With the abundant profits brought by the market monopoly and the strategic use of intellectual property rights (IPRs), patent management has played an important role in the effective operation of enterprises. For instance, the royalty income of IBM has topped a billon US dollars every year, which is approximately at least one ninth of its annual gross profit before tax. The success of Texas Instrument's patent in court enabled the firm to earn higher royalty payments from other firms in the semiconductor industry (Rivitte & Kline, 2000). Thus, IPRs have become important company assets, and patent management has played a pivotal role in corporate management and performance.

The current studies on patent management apply the International Patent Code (IPC) or the United States Patent Code (UPC) to identify patents. Ernst (1997) defines the patents with IPC classification code G05B019/00-G05B019/417 as CNC-technology in the machine tool industry to forecast the diffusion of CNC-technology. Narin, Noma, and Perry (1987) use 15 categories of UPC codes, e.g. code 260, 424, etc., to examine the links between corporate patents and indicators of pharmaceutical corporate performance. However, in terms of patent management, the IPC or the UPC system is too general to satisfy the needs for technological forecasting, research planning, technological positioning or strategy making (Archibugi & Pianta, 1996). The result of the analysis by the above two systems is insufficient to reflect the technological niche of a company and mis-categorization results in further difficulties for patent management. In addition, both the UPC and the IPC system are static systems, which mean they do not evolve with the development of technologies. Thus, the main objective of this study is to propose a patent classification system tailored to the needs of a specific industry for the management of its patents.

This study applies citation approach in bibliometrics to create a patent classification system based on the following grounds: First, dissertations and patents are both instruments that record the results of research. In addition, published dissertations and patent specifications are required to identify their cited documents and patents. Secondly, in bibliometrics, the use of citation approach for the assessment of similarity for the classification of documents is a mature methodology. Given the above, it is feasible to apply citation analysis of bibliometrics to patent classification.

In bibliometrics, there have been extensive studies on the assessment of document similarities. For instance, Kessler (1963) proposed the approach of bibliographic coupling, and Small (1973) proposed the co-citation approach. For bibliographic coupling, citing documents are the subject of the analysis. The degree of bibliographic coupling for documents A and B is reflected in the frequency of the documents that are co-cited by both A and B. The focus of the co-citation analysis is on the documents cited, by calculating the frequency of A and B that are co-cited by specific documents. They assess the similarity of A and B based respectively on the number of co-citing or co-cited documents, as illustrated in Fig. 1. In bibliometrics, the perception of being co-cited is applied to evaluating the similarities in documents because the number of documents co-cited by A and B is limited to that of the reference documents but the number of A and B being

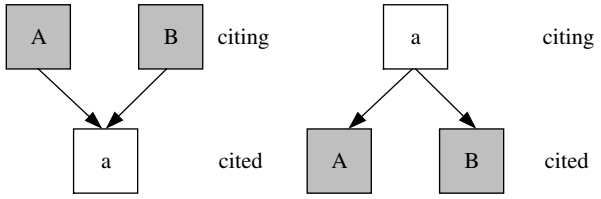

Fig. 1. Bibliographic coupling vs. co-citation.

co-cited is not a subject to this limitation. Therefore, this study will use the co-cited frequency of patents to assess the similarities in patents and to create a patent classification system.

The purpose for the assessment of document similarities is to classify documents. The classified levels include the document itself, its author, and the journal that contains the document; each with a different application. For instance, documents co-citation is used to conduct searches on similar documents (Akin, 1998). Journal co-citation is of interest to the collection manager concerned with developing core journal lists, selecting journals and evaluating collections that serve particular research-oriented constituencies (Holsapple, Johnson, Manakyan, & Tanner, 1995; McCain, 1991). Author co-citation analysis has been used in analyzing the intellectual structure of science studies (Culnan, 1987; Culnan, O'Reilly, & Chatman, 1990; Eom, 1996; Hoffman & Holbrook, 1993; McCain, 1990; White & Griffith, 1981). The application of the approach of document classification to the relevant research on patents may serve different purposes, but citation analysis is rarely used for patent analysis. Stuart and Podoly (1996) applied the conception of corporate patent co-citation to enable firms to be positioned and grouped according to the similarities in their patents, which was a pioneer application of this approach to the patent management.

Inappropriate or erroneous citation will endanger the result of co-citation research. Because co-citation is a noisy measure of similarity, many scholars have used other approaches to verify the applicability of the result of the analysis (Hayes, 1983; McCain, 1986). Similarly, an inappropriate patent citation will have adverse effects on the quality of a patent classification system. As a result, prior using patent cited material, evaluating the probabilities of erroneous patent citation could help users understand the appropriateness of this approach.

Reference documents of academic papers and patents cited by patent specifications manifest the inheritance of research projects and the contributions of researchers. Accumulation and uniqueness are two essential elements for academic research and invention, and researchers are rewarded by the first publication of their works. The motives and the accuracy of their document citation do not affect the authors' achievements. In contrast, the exclusive right of a patent owner lies in the patent claims and the economic value generated thereby (Dasgupta & David, 1987). Citations serve to show how the claimed invention differs from the "prior art". The basic purpose of citing "prior art" in patent files is to inform the patent owner and the public in general that such patents or printed publications are in existence and should be considered when evaluating the validity of the patent claims (USPTO, 2001). Patent applicants and examiners are cautious with patent citations, which help define a patent and have direct influence on its economic value. Given the above, we may conclude that the noisy disturbance of patent citation is less than that of cited

documents for academic papers. Thus, the analysis of patent citation may yield more credible results.

Hence, this study uses the co-citation analysis, which is applicable to patents, to propose an approach called the patent co-citation approach (PCA) to create a patent classification system. This study has the following two contributions. First, we establish a patent classification system that reflects the similarities in patents. The system overcomes the flaws of the IPC and the UPC in that they are too general to perform the patent analysis for a specific industry. Secondly, the patent classification system gives patent managers a clearer picture of basic patents for a specific industry. Lastly, the classification system reveals the relationship among categories of technologies and the evolution of a technology category.

The results of this study can be applied to research planning, patent value assessment, the composition of patent portfolio, and the making of licensing strategy. The remainder of the paper is organized as follows: Section 2 proposes the patent co-citation approach. Section 3 verifies applicability of PCA by creating a patent classification system for the semiconductor foundry industry. Sections 4 and 5 provide discussion and conclusions, respectively.

2. The patent co-citation analysis

The PCA is a methodology for creating a patent classification system by classifying industry *basic patents*. After the patent classification system is built, the *target patents* will be classified by being compared with the basic patents. In this research, target patents are patents to be classified and basic patents are patents repeatedly cited by target patents. The conception and the application of the PCA are illustrated in Fig. 2.

For instance, Q_1–Q_7 are target patents, from which basic patents P_1–P_5 are selected. The directional line between the target patents and the basic patents indicates the referential relationship between these two groups. According to the similarities in basic patents, two technology categories F_1–F_2 are defined. P_1 and P_2 are covered by the F_1 category; P_3, P_4, and P_5 are assigned to the F_2 category. If when Q_1 cites P_1, Q_1 is classified to be the F_1 technology.

To complete this classification system, the analysis is divided into three phases, as shown in Fig. 3. Phase I selects the proper database to query target patents and specify basic patents. Phase II uses the co-cited frequency of the basic patent pairs to assess the similarity. Phase III uses factor

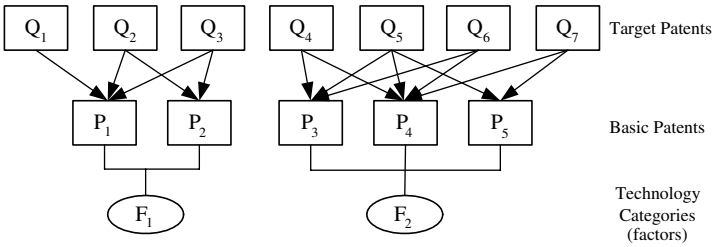

Fig. 2. The conception and the application of PCA.

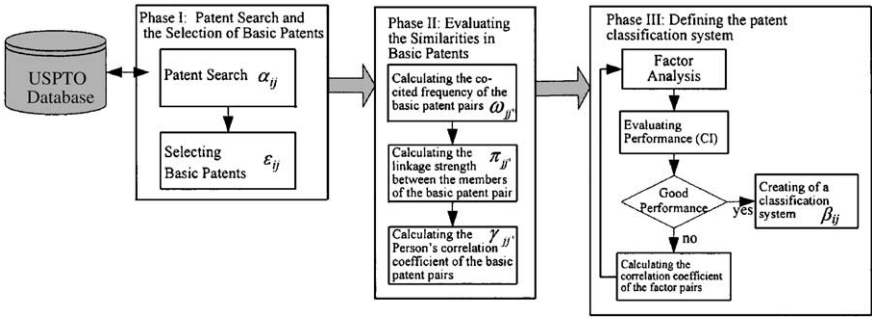

Fig. 3. The analysis process of the PCA.

analysis to group basic patents into a smaller set of factors and to evaluate the performance of the PCA. Details of the above three phases are provided below.

2.1. Phase I: searching for patents and defining industry basic patents

In this phase, the proper database will be selected to conduct the patent search, and basic patents will be specified from the search results.

2.1.1. Patent search

Patent offices world-wide, such as the United States Patent Trademark Office (USPTO), the European Patent Office (EPO), and the Japan Patent Office (JPO), are dedicated to the establishment of patent databases to improve the dispersion of patent information. The database of the USPTO is one of the favored sources to conduct a patent search because the US market is an important market for technology-transfer and international trade, which is combined with the territoriality of patent protection, luring inventors to file patent applications in the US.

After a database is picked, the next step is to select patent owners according to the objective of the patent management. For instance, provided that the Taiwan Semiconductor Manufacturing Company (TSMC) develops a patent classification system, adaptive to the semiconductor foundry industry, the scope of the patent search will cover the patents of the TSMC and those of its competitors, such as the United Microelectronics Company (UMC) and the Chartered Semiconductor Manufacturing Corporate (Chartered).

The selected patents will be divided into two groups: *target patents* and *candidate of basic patents*. We denote Q_i as target patent i and CP_j as a candidate for basic patent j, respectively. Target patents are citing patents to be classified. Candidates of basic patents are those patents that are cited by target patent. The relationship between target patents and candidate of basic patents are demonstrated in Fig. 4.

The matrix shown in Eq. (1) describes the referential relationship between each pair of the target patents and candidate of basic patents, where M is the number of target patents, and N is the number of candidate of basic patents.

$$[\alpha_{ij}]_{M\times N}, \quad \text{where } \alpha_{ij} = \begin{cases} 1 & Q_i \text{ cites } \mathrm{CP}_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

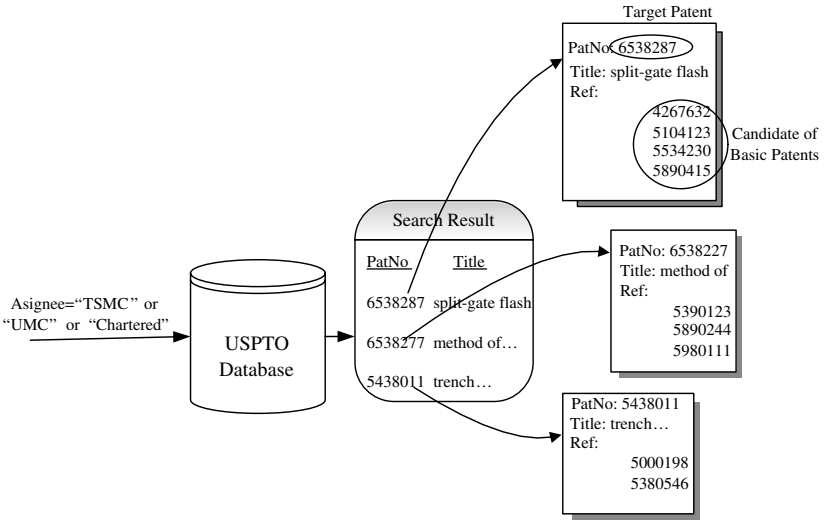

Fig. 4. The definition of target patents and candidate of basic patents.

2.1.2. The selection of basic patents

The PCA defines technology categories with industry basic patents. The so-called basic patents here are the patents being repeatedly cited by later patents. Reasons to define the basic patent by its cited frequency are twofold:

- The more a specific early patent is cited by later patents, the more likely it is to be the foundation of these later patents (Mogee, 1997); and
- The cited frequency of a patent is an important indicator to evaluate the quality of a patent (Hall, Jaffe, & Trajtenberg, 2000; Harhoff, Narin, Scherer, & Vopel, 1999; Narin et al., 1987; Trajtenberg, 1990).

Hence, this study uses the cited frequency of a candidate basic patent to select basic patents. The frequency of CP_j being cited is demonstrated in Eq. (2).

$$CS_j = \sum_{i=1}^{M} \alpha_{ij}, \quad 1 \leqslant j \leqslant N \tag{2}$$

In this study, CP_j become a basic patent if CS_j is greater than or equal to the threshold c for selecting basic patents. The threshold c is determined by the patent manager by comparing classifying performances at different threshold values.

After identifying basic patents, a new matrix, as Eq. (3), can be created from the relationship between the basic patents and the target patents. We denote P_j as basic patent j, n as the number of basic patents, and m as the number of target patents which can be classified by basic patents.

$$[\varepsilon_{ij}]_{m\times n}, \quad \text{where } \varepsilon_{ij} = \begin{cases} 1 & Q_i \text{ cites } P_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

2.2. Phase II: assessment of the similarities in basic patent pairs

In this paper, the Pearson correlation coefficient is employed to assess the similarity for a basic patent pair. Three steps are required to obtain the similarity for each basic patent pair:

Step 1: Calculate the co-cited frequency of each basic patent pair;
Step 2: Calculate the linkage strength of each basic patent pair;
Step 3: Calculate the Pearson correlation coefficient of each basic patent pair.

The details of each step are given below.

Step 1: Calculate the co-cited frequency of each basic patent pair.

Given patents j and j', the co-cited frequency of the two patent is

$$\omega_{jj'} = \begin{cases} \sum_{i=1}^{m} \varepsilon_{ij}\varepsilon_{ij'} & \text{if } j \neq j' \\ 0 & \text{if } j = j' \end{cases} \quad 1 \leqslant j \leqslant n, \;\; 1 \leqslant j' \leqslant n \tag{4}$$

where ε_{ij} and $\varepsilon_{ij'}$ are citing relationships defined as Eq. (3).

A symmetrical matrix $[\omega_{jj'}]_{n\times n}$ can be obtained in this step after calculating all of the co-cited frequencies of n basic patents.

Step 2: Calculate the linkage strength of each basic patent pair.

The linkage strength of a basic patent pair is calculated as following:

$$\pi_{jj'} = \begin{cases} \dfrac{\omega_{jj'}}{S_j + S_{j'} - \omega_{jj'}} & \text{if } j \neq j' \\ 0 & \text{if } j = j' \end{cases} \quad 1 \leqslant j \leqslant n, \;\; 1 \leqslant j' \leqslant n \tag{5}$$

where $\omega_{jj'}$ is the co-cited frequency calculated in the previous step; $S_j = \sum_{i=1}^{m} \varepsilon_{ij}$ is the cited frequency of a basic patent j.

The linkage strengths of basic patent pairs form a new symmetrical matrix $[\pi_{jj'}]_{n\times n}$, which is the input of the next step.

Step 3: Calculate the Pearson correlation coefficient of each basic patent pair.

Given basic patents j and j', the steps to calculate the Pearson correlation coefficient, $\pi_{jj'}$, of a basic patent pairs are as following:

Step 3.1: Divide the linkage strengths of pairs of basic patents into two groups.

The first group is $\Pi_j = \{\pi_{kj}, k \neq j, j'\}$ *and the other is* $\Pi_{j'} = \{\pi_{kj'}, k \neq j, j'\}$.

Step 3.2: Calculate the Pearson correlation coefficient by the following equation:

$$r_{jj'} = \begin{cases} \dfrac{(n-2)\sum_{k=1}^{n} \pi_{kj}\pi_{kj'} - \sum_{k=1}^{n} \pi_{kj} \sum_{k=1}^{n} \pi_{kj'}}{\sqrt{(n-2)\left(\sum_{k=1}^{n} \pi_{kj}^2\right) - \left(\sum_{k=1}^{n} \pi_{kj}\right)^2}\sqrt{(n-2)\left(\sum_{k=1}^{n} \pi_{kj'}^2\right) - \left(\sum_{k=1}^{n} \pi_{kj'}\right)^2}} & \text{if } j \neq j' \\ 1 & \text{if } j = j' \end{cases} \tag{6}$$

where $\pi_{kj} \in \Pi_j$ is the linkage strength between basic patent j and k; $\pi_{kj'} \in \Pi_{j'}$ is the linkage strength between basic patent j' and k.

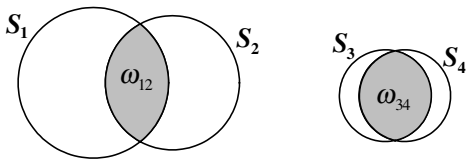

Fig. 5. The flaws in the use of the frequency of co-citation on the assessment of similarities in basic patent pairs.

After finishing this step, we have a matrix of Pearson correlation coefficient of basic patents, $[\gamma_{jj'}]_{n\times n}$, to measure similarities between basic patents.

The main reason for employing Eq. (6) to measure the similarity between basic patents is that this equation overcomes the problem caused by using only the co-cited frequencies or the linkage strengths between basic patents.

The problem of over-estimating or under-estimating similarities in basic patent pairs occurs when singly employing the co-cited frequency as the indicator of the similarity between two basic patents. Consider the following example. There are four basic patents, as shown in Fig. 5. The areas of circles S_1, S_2, S_3 and S_4 represent the cited frequencies for P_1, P_2, P_3 and P_4, respectively. Also, the gray oval areas ω_{12} and ω_{34} represent the co-cited frequencies for the two basic patent pairs (P_1 and P_2, P_3 and P_4), respectively. The two pairs of basic patents have the same similarity when ω_{12} equals to ω_{34}. However, the similarities in the basic patent pair P_3 and P_4 should be greater than those in P_1 and P_2 because the cited frequencies of P_3 and P_4 are smaller than that of P_1 and P_2.

A bias is caused when only employing linkage strength to measure the similarity of a basic patent pair, Yulan and Cheung (2002). The linkage strength is the direct linkage between the basic patent pair. However, high linkage strength might be randomly generated. A better method is to measure the consistency in linkage strengths of each patent in the same pair to other basic patents. In other words, P_j and $P_{j'}$ are highly similar to each other when both P_j and $P_{j'}$ have high (or low) linkage strength with other basic patents, except for themselves.

Using the Pearson correlation coefficient has two advantages in measuring the similarity of a basic patent pair. Firstly, the Pearson correlation coefficient functions as a measure, not just of how often that pair of basic patents were co-cited, but of how similar their linkage strength profiles are. Secondly, the Pearson correlation coefficient standardizes the linkage strength, which may solve the problem caused by scale (McCain, 1990).

2.3. Phase III: creation of a patent classification system

The bibliometrics generally employs factor analysis, cluster analysis, or multi-dimensional scaling to classify documents, journals, and authors. This study adopts factor analysis based on two considerations. Firstly, the loading of the patent (variables) on the technology category (factor) indicates the degree of importance for the basic patent to the technology category. Secondly, if necessary, factor analysis may be repeated to create a hierarchical classification system.

2.3.1. Factor analysis

The input for factor analysis is the Pearson correlation coefficient of the basic patents; and after the factor analysis, we will have G technology categories. Then, for those ε_{ij} equal to 1, we modify the value of ε_{ij} to the g that is the technology category of basic patent P_j. Thereby, we create a new matrix $[\beta_{ij}]_{m\times n}$ that indicates both the technology category that a basic patent belongs to and the referential relationships between the target patents and the basic patents, as shown in Eq. (7).

$$[\beta_{ij}]_{m\times n} \quad \text{where } \beta_{ij} = \begin{cases} g & Q_i \text{ cites } P_j \\ 0 & \text{otherwise} \end{cases} \quad 1 \leqslant g \leqslant G \tag{7}$$

2.3.2. Performance evaluation

The feasibility of the classification system generated by PCA can be evaluated from three aspects: the ease of naming the technology category, the fitness between the patent classification system and industry technologies, and the consistency of the classifying result. The easier the category-naming is, the more features are shared by the basic patents that are assigned to the same category. Moreover, a good patent classification system can describe the features of an industry's technology in an appropriate way. High consistency of the classifying result indicates good performance of the classification system.

The former two evaluation approaches must be carried out by industry experts, which are qualitative evaluation indicators. The consistency index CI, the third approach, can be obtained by the following equation:

$$\mathrm{CI} = \frac{m - x}{m} \tag{8}$$

where x is the number of target patents that are multiply classified.

The more multiple classifications that occur, the worse the performance of the patent classification. In the PCA, multiple classifications for a target patent are possible. The category for the target patent is decided by the category of its cited basic patents. Also, basic patents are grouped into several technological categories. When basic patents cited by a target patent do not belong to the same technology category, the target patent has multiple classifications. For instance, the target patent Q_1 cites two basic patents, P_1 and P_2. When P_1 and P_2 do not belong to the same technology category, Q_1 has duplicated classifications.

2.4. Illustration of the PCA model

In order to demonstrate the conception of the PCA and its applicability, an example is given below to explain the analytical process and the employment of the classification system.

Company A uses the PCA to create a patent classification system. One of its researchers conducts a patent search on the database of the USPTO for its own patents and for those of its competitors (target patents) and for the patents which are cited by the target patents (candidate of basic patents). The citation network is illustrated in Fig. 6 and can be demonstrated by the matrix $[\alpha_{ij}]_{8\times 10}$. Here C, D, E, G, I, K, L, and M represent target patents; while A, B, C, D, E, F, G, H, I, and J stand for candidate basic patents. The number represents the filing order of patents. The directional lines represent the referential relationship between the target patents and the candidate of basic patents.

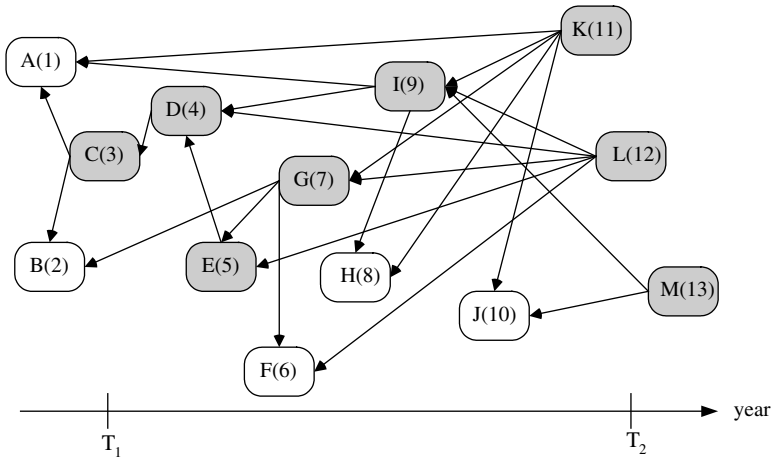

Fig. 6. The citation network of the patents.

2.4.1. Phase I: patent searches and the selection of basic patents

All the cited patents by target patents are candidates for basic patents. However, the selection of basic patents is based on the frequency of being referred to by the target patents. The steps taken in this process are as follows: Set the threshold for the selection of basic patents at 1, eliminate C, which is a cited patent failing to be selected as a basic patent, eradicate the patent D which refers to only C. The result of the process is demonstrated by the matrix $[\varepsilon_{ij}]_{7\times 9}$ as shown in Table 1.

2.4.2. Phase II: assessment of the similarities in basic patents

The Pearson correlation coefficient matrix $[y_{jj'}]_{9\times 9}$ for basic patents is a final output from phase II, and the result thereof is shown in Table 2.

2.4.3. Phase III: creation of a patent classification system

In this phase, factor analysis is applied to the classifying of basic patents in order to create a patent classification system. This study uses principal component analysis with the promax rotation to extract factors. The eigenvalue greater than 1 criteria is applied to obtain 2 factors. The result of the analysis divides the patents into two categories. The loading of the basic patents

Table 1
The citation matrix between target patents and basic patents

ε_{ij}	P_1 (A)	P_2 (B)	P_3 (D)	P_4 (E)	P_5 (F)	P_6 (G)	P_7 (H)	P_8 (I)	P_9 (J)
Q_1 (C)	1	1	0	0	0	0	0	0	0
Q_2 (E)	0	0	1	0	0	0	0	0	0
Q_3 (G)	0	1	0	1	1	0	0	0	0
Q_4 (I)	1	0	1	0	0	0	1	0	0
Q_5 (K)	1	0	0	0	0	1	1	1	1
Q_6 (L)	0	0	1	1	1	1	0	1	0
Q_7 (M)	0	0	0	0	0	0	0	1	1

Table 2
The Pearson correlation coefficient of basic patents

$\gamma_{jj'}$	P_1 (A)	P_2 (B)	P_3 (D)	P_4 (E)	P_5 (F)	P_6 (G)	P_7 (H)	P_8 (I)	P_9 (J)
P_1	1								
P_2	−0.69	1							
P_3	−0.05	0.367	1						
P_4	−0.73	0.587	0.28	1					
P_5	−0.73	0.587	0.28	1	1				
P_6	−0.06	−0.3	0.475	−0.01	−0.01	1			
P_7	0.685	−0.32	0.039	−0.72	−0.72	0.148	1		
P_8	0.038	−0.51	0.015	−0.17	−0.17	0.658	0.325	1	
P_9	0.399	−0.59	0.244	−0.42	−0.42	0.764	0.469	0.591	1

(variables) on these two categories decides the category for the basic patents. As a result, the basic patents, P_1, P_2, P_3, P_4, P_5, and P_7 are assigned to the first category. The basic patents, P_6, P_8 and P_9 are allocated to the second category. In the PCA, the interpretation or the definition of each factor is based on those patents with high loadings. Only patents with loadings greater than 0.7 are likely to be useful in interpreting the factor. The results of the analysis are shown in Table 3.

As shown in Eq. (7), Table 3 is adjusted according to the results of the factor analysis, thereby creating a patent classification system $[\beta_{ij}]_{7\times 9}$ as shown in Table 4. In Table 4, the target patent Q_1 cites the basic patents P_1 and P_2 respectively. Since P_1 and P_2 belong to the first category, Q_1 is assigned to the same category as that of P_1 and P_2. Q_5 cites the basic patents, P_1, P_6, P_7, P_8 and P_9, among which P_1 and P_7 belong to the first category, while P_6, P_8 and P_9 belong to the second category. As a result, the classification of Q_5 is duplicated. In practical application, the target patent subject to multiple classifications could be assigned to the category with higher citing frequency. For instance, the frequency for Q_5 to cite the patents in the first and the second category are 2/5 and 3/5, respectively. As a result, Q_5 is assigned to the second category.

Among seven target patents, Q_5 and Q_6 are subject to multiple classifications. The consistency rate (CI) for the result of classification is 5/7.

Table 3
The structure loading of factors

Patent(variable)	Factor 1	Factor 2
P_4 (E)	**0.961**	−0.525
P_5 (F)	**0.961**	−0.525
P_1 (A)	**−0.950**	0.442
P_7 (H)	**−0.940**	0.511
P_2 (B)	**0.853**	−0.712
P_3 (D)	**0.630**	0.137
P_6 (G)	−0.147	**0.976**
P_9 (J)	−0.671	**0.916**
P_8 (I)	−0.481	**0.896**
Eigenvalues	6.388	1.664
% variance	70.874	18.487
Cumulative%	70.874	89.461

Table 4
The classification system of patents

β_{ij}	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
Q_1	1	1	0	0	0	0	0	0	0
Q_2	0	0	1	0	0	0	0	0	0
Q_3	0	1	0	1	1	0	0	0	0
Q_4	1	0	1	0	0	0	1	0	0
Q_5	1	0	0	0	0	2	1	2	2
Q_6	0	0	1	1	1	2	0	2	0
Q_7	0	0	0	0	0	0	0	2	2

2.5. The features of the PCA

Compared to static systems, such as the IPC or the UPC, the PCA is a more dynamic and self-organized methodology that reflects the technological status of an industry. However, technologies continue to develop, so a system should keep pace with the evolution of a technology. The PCA may, through the use of computer systems facilitate the updating of a patent classification system and the re-classification of patents, and thus reduce the cost for patent management.

3. Experiment and results

In order to evaluate the feasibility of the PCA, here we pick the patent-crowded semiconductor foundry industry to conduct an analysis of the proposed classification system. In the high-tech industry, corporation's competitive edge originates from their technological capabilities. Enterprises with a higher market share usually take the lead in the industry technology. Since TSMC, UMC, IBM and Chartered are leading manufacturers in the semiconductor foundry market in 2002 (Semico Research), we choose the patents held by these four companies and issued by the USPTO to conduct an analysis.

TSMC, UMC, and Chartered are engaged only in the foundry manufacturing of semiconductors. Their patents are expected to be related to the semiconductor technologies and therefore can all be used by the analysis. TSMC holds 2272 patents, UMC 2324, and Chartered 550. IBM is a highly diversified corporation, and semiconductor manufacturing is only one of its business branches. As a result, we only choose its semiconductor-related patents for the analysis. Detailed information on these four companies is shown in Table 5.

3.1. Phase I: patent searches and the selection of basic patents

These four companies hold a total of 8967 patents (target patents) and they cite 36 795 patents (candidates for basic patents). The referential relationship between the target patents and the candidates for basic patents form a sparse matrix $[\alpha_{ij}]_{8967\times 36\,795}$. In the next step, we use the cited frequency of the candidate of basic patents to select the basic patents of the semiconductor foundry industry. In this process, we find that older patents tend to have higher cited frequency because they have been available longer than for the more recent patents. As a result, newer

Table 5
Profile of the analyzed companies

	TSMC	UMC	IBM	Chartered
Year for the establishment	1987[a]	1970[b]	1911[c]	1987[d]
Market share in 2002	41.5%	17.4%	6.2%	4.3%
The number of the patents[e]	2272	2324	3821	550

[a] http://www.tsmc.com.
[b] http://www.umc.com.
[c] http://www.ibm.com.
[d] http://www.charteredsemi.com.
[e] Issued between 1991/1/1 and 2003/4/30.

patents are less likely to be chosen as basic patents, which in turn, indicate that new technological development would be left out of this analysis. In order to address this problem, Eq. (2) is adjusted and transformed into Eq. (9).

$$\mathrm{ST}_j = \sum_{i=1}^{M} \alpha_{ij} \times \mathrm{wt}_i \tag{9}$$

where wt_i is the weight of the target patent i, which is obtained by subtracting 1985 from the apply year of the target patent.

Using $\mathrm{ST}_j \geqslant 174$ as the criteria to select basic patents, 240 basic patents are selected from the 36 795 candidate of basic patents. In turn, these 240 patents are used to create a classification system for the semiconductor foundry industry. We find that 2120 out of a total of 8967 target patents refer to basic patents. The referential relationship between the basic patents and the target patents can be demonstrated by the matrix $[\varepsilon_{ij}]_{2120\times240}$.

3.2. Phase II: the evaluation of the patent similarities

For the first step, calculate the co-cited frequency of C_2^{240} basic patent pairs by 2120 target patents with Eq. (4), and the result is shown in the matrix $[\omega_{jj'}]_{240\times240}$. The matrix $[\omega_{jj'}]_{240\times240}$ is taken into Eq. (5) to obtain the linkage strength between basic patent pairs, and the result is demonstrated in the matrix $[\pi_{jj'}]_{240\times240}$. The matrix $[\pi_{jj'}]_{240\times240}$ is taken into Eq. (6) and the correlation coefficient matrix $[\gamma_{jj'}]_{240\times240}$ for the basic patent pairs is derived. In the next step of factor analysis, we analyze the correlation coefficient matrix to obtain the configuration of the basic patents.

3.3. Phase III: the creation of the classification system

The matrix of basic patent pairs correlation coefficient $[\gamma_{jj'}]_{240\times240}$ is factor-analyzed using a principal components analysis with promax rotation. Based on eigenvalue more than 1 criterion, 31 factors are retained. These 31 factors account for 91.88% variance. The result is represented in Table 6. The marginal variance explained by the 17th factor is low. Thus, 16 factors are retained, which account for 82.5% of the variance.

Table 6
Eigenvalues and variances explained by factors

Factor	Eigenvalues	Variance explained %	Cumulative variance %
1	54.271	22.613	22.613
2	35.016	14.590	37.203
3	19.977	8.324	45.527
4	13.157	5.482	51.009
5	9.869	4.112	55.121
6	9.804	4.085	59.206
7	8.731	3.638	62.844
8	7.931	3.305	66.149
9	6.561	2.734	68.882
10	6.449	2.687	71.569
11	6.046	2.519	74.089
12	5.538	2.307	76.396
13	4.740	1.975	78.371
14	3.820	1.592	79.963
15	3.280	1.367	81.330
16	2.840	1.183	82.513
17	2.612	1.088	83.601
18	2.071	0.863	84.464
19	1.938	0.808	85.272
20	1.862	0.776	86.047
21	1.643	0.685	86.732
22	1.621	0.675	87.407
23	1.425	0.594	88.001
24	1.413	0.589	88.590
25	1.300	0.542	89.131
26	1.176	0.490	89.621
27	1.164	0.485	90.106
28	1.102	0.459	90.566
29	1.081	0.450	91.016
30	1.063	0.443	91.459
31	1.028	0.428	91.887

Among 240 basic patents, 18 have high loading on dropped factor and 24 patents have duplicated loading on factors 1–16. Because these 42 patents cannot be distinctively classified, the target patents that cite these patents also cannot be classified. The target patents that can be classified are reduced from 2140 to 1673 and the matrix $[\varepsilon_{ij}]_{2140\times240}$ is diminished as $[\varepsilon_{ij}]_{1643\times198}$. Finally, the result of classifying 198 basic patents is used to adjust the matrix $[\varepsilon_{ij}]_{1643\times198}$ and the matrix $[\beta_{ij}]_{1643\times198}$ for the patent classification system of the semiconductor industry is created. After analyzing $[\beta_{ij}]_{1643\times198}$, we find that 256 out of 1643 basic patents are subject to multiple classifications. The performance indicator CI for evaluating the consistency on classification is 0.844.

After the first factor analysis, the patents related to the semiconductor foundry industry are divided into 16 technology categories (A_1–C_7). In order to understand the relationship between technology categories, to provide patent classification more flexibility, and to reduce multiple classifications, the second-order factor analysis is conducted. As a result, the said 16 categories are

Table 7
The result of the patent classify on the semiconductor foundry industry

Category	Sub-category	Factor	Name	The number of basic patents	The number of target patents
A			*Front end process*	**69**	**485**
	A_1	1	DRAM and SRAM	35	87
	A_2	5	STI	12	152
	A_3	10	STI related process	10	141
	A_4	11	Lithographic	6	42
	A_5	12	Other memory	6	63
B			*Back end process*	**61**	**711**
	B_1	2	Metal interconnection	29	284
	B_2	4	Dual Damascene	17	194
	B_3	8	Contact	9	165
	B_4	13	Air gap (air bridge)	6	68
C			*Process integration*	**68**	**736**
	C_1	3	Process for deep trench	26	238
	C_2	6	Process for integration	10	112
	C_3	7	Process for high density package	9	78
	C_4	9	SOI	8	116
	C_5	14	MOS structure	5	73
	C_6	15	Process for fabricating MOS	5	65
	C_7	16	Related process for the gate of MOS	4	54

$CI = 0.963$, $CI = 0.844$

grouped into A, B and C, three upper-level categories. When classifying 1643 patents, 61 patents are duplicated and allocated to the upper-level categories A, B and C. On the upper level classification system, the performance indicator for classification consistency is raised to $CI = 0.963$.

After the factor analysis was completed, electronics professors teaching in KunSung University of Technology and senior engineers working for Taiwan Applied Materials Inc. were invited to name the three categories and 16 sub-categories of hierarchical classification system. The results of the naming and the number of basic patents and target patents are shown in Table 7.

4. Discussion

This section will discuss the issues on application of PCA. The PCA approach is more suitable for patent-crowded industries, such as the semiconductor industry and the electronics industry. Crowding of patents indicates that innovation is incremental but not radical. As a result, every application for a patent must inevitably cite several "prior arts". The citing relationship among patents will form a referential network. The PCA approach is a study of this network to establish a patent classification system.

Though the PCA approach may solve the problems caused by the IPC or the UPC, this approach still has its limitations. The application of this approach may give rise to problems such as multiple classifications or non-classification. Multiple classifications are caused by the citing of the

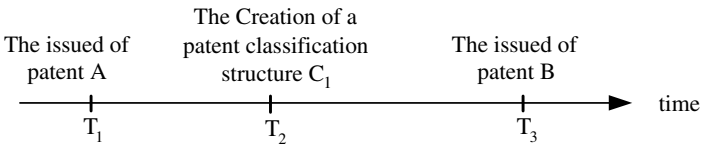

Fig. 7. The relationship between the issue time of target patents and the creation of a patent classification system.

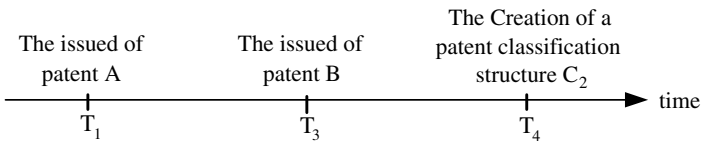

Fig. 8. The relationship between the issued time of target patents and the creation of a patent classification system.

basic patents from highly relevant technological categories. One possible solution for this problem is to use the hierarchical classification system to reduce the chances for duplication. The hierarchical classification system is created by repetitive application of factor analysis, applying the correlation coefficient derived from the last factor analysis. In this structure, a category of the upper level is derived from the integration of the relevant lower level categories.

Non-classification is due to the fact that the target patents do not cite the basic patents defined in this study. However, non-classification has positive effects on patent management, and below, we discuss the effects of non-classification in two scenarios, as shown in Fig. 7. In the first scenario, target patent A was issued before the creation of the patent classification system C_1, which indicates that patent A did not cite basic patents, and which also implies that patent A is not itself a main stream technology. As a result, ignoring patent A will not have a visible effect on the performance of patent management. In the second scenario, target patent B cannot be classified with C_1, which may imply that patent B is not covered by main stream technology, or may reveal the development of new technologies. At this point, a new patent classification system should be developed to accommodate new technologies or the performance of patent management will be compromised.

With the creation of C_2, the patent B, which originally could not be classified, can possibly be classified under the C_2 structure. Consequently, in order to maintain the classification performance, basic patents and the patent classification system should be updated regularly.

In this paragraph, we further develop the first scenario to explore its implications for patent management when the target patents A and B cannot be classified with C_2, as shown in Fig. 8. First of all, in terms of main-stream technologies for industries, patent A is less likely to become a main-stream technology than patent B because patent A is older than patent B. Secondly, in terms of the duration of patents, patent A has a shorter remaining term than patent B. Since the duration of a patent is one the important elements that decides the value of a patent, the economic value of the patent A is less than that of the patent B. Given the above, the target patents that could not be classified should be abandoned to reduce maintenance fees, except for those issued more recently, which should be monitored for their development and new applications.

In light of the above, we may conclude that the patent classification system should be updated regularly. The frequency for updating differs with each industry. In addition, for some of the older patents that cannot be classified, one may consider abandoning this kind of patent to reduce maintenance cost.

5. Conclusion

In order to overcome the flaws of the IPC or the UPC system, this study applies the co-citation analysis used in bibliometrics to propose a methodology for establishing a patent classification system. This approach is composed of three parts: selecting basic patents, assessing the similarities of the basic patents, and establishing a patent classification system. To give a clearer picture on the conception of the PCA approach, this study demonstrates the analytical process and the practical applications of this approach by using a set of simulation materials to demonstrate the concepts of the PCA. In order to further verify the feasibility of the PCA, this approach has been tested on the patents held by four leading semiconductor foundry manufacturers, TSMC, UMC, IBM, and Chartered, in order to establish a patent classification system specifically for this industry. In this system, there are three categories and these three categories are divided into 16 sub-categories, which were subsequently determined by a professor and a senior engineer from the industry. These experts' views on the naming of each category are quite consistent. The rate on the consistency of the classifying result for the categories and sub-categories reached 84.4% and 96.3%, respectively. Both the qualitative evaluation by the experts and the quantitative indicator demonstrate the feasibility of the PCA. Further, the PCA is suitable for a patent-crowded industry, such as the semiconductor industry or the electronics industry. Using this approach for an industry which has sparse patents could be less effective and the results would not be workable for patent management.

In the analytical process of the semiconductor foundry industry, we discovered that the PCA approach is subject to two major problems: multiple classifications and non-classification and this study discusses their effects on the performance of the patent management and provides solution thereto. Future research will use the classification system for research planning and the analysis of patent portfolio and technology positions in industry, so as to provide more applicable information for the industry.

Acknowledgement

The research on which this article is based was support by the NSC (92-2416-H-224-004). This support is gratefully acknowledged.

References

Akin, L. (1998). Methods for examining small literatures: explication, physical analysis, and citation patterns. *Library and Information Science Research, 20*(3), 251–270.

Archibugi, D., & Pianta, M. (1996). Measuring technological change through patents and innovation survey. *Technovation, 16*(9), 451–468.

Culnan, M. J. (1987). Mapping the intellectual structure of MIS, 1980–1985: a co-citation analysis. *MIS Quarterly, 11*(3), 341–352.

Culnan, M. J., O'Reilly, C. A., & Chatman, J. A. (1990). Intellectual structure of research in organizational behavior, 1972–1984: a co-citation analysis. *Journal of American Society for Information Science, 41*(6), 453–458.

Dasgupta, P., & David, P. A. (1987). Information disclosure and the economics of science and technology. In R. George (Ed.), *Arrow and the ascent of modern economic Theory*. Feiwel Basingstoke, Hampshire: Macmillan.

Eom, S. B. (1996). Mapping the intellectual structure of research in decision support systems through author co-citation analysis (1971–1993). *Decision Support System, 16*, 315–338.

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

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2000). Market value and patent citation: a first look. NBER Working Paper W7435.

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented innovation. *The Review of Economics and Statistics, 81*(3), 511–515.

Hayes, R. M. (1983). Citation statistics as a measure of faculty research productivity. *Journal of Education Librarianship, 23*(3), 151–172.

Hoffman, D. L., & Holbrook, M. B. (1993). The intellectual structure of consumer research: a bibliometrics study of author co-citations in the first 15 years of the journal of consumer research. *Journal of Consumer Research, 19*, 505–517.

Holsapple, C. W., Johnson, L. E., Manakyan, H., & Tanner, J. (1995). An empirical assessment and categorization of journals relevant to DSS research. *Decision Support Systems, 14*(4), 359–367.

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

Kessler, M. M. (1963). An experimental study of bibliographic coupling between technical papers. *IEEE Transaction on Information Theory, 49*.

McCain, K. W. (1986). Co-cited author mapping as a valid representation of intellectual structure. *Journal of American Society for Information Science, 37*(3), 111–122.

McCain, K. W. (1990). Mapping authors in intellectual space: a technical overview. *Journal of American Society for Information Science, 41*(6), 433–443.

McCain, K. W. (1991). Mapping economic through the journal literature: an experiment in journal co-citation analysis. *Journal of American Society for Information Science, 42*(4), 290–296.

Mogee, M. E. (1997). Patent analysis methods in support of licensing. *Technology Transfer Society Annual Conference*.

Narin, F., Noma, E., & Perry, R. (1987). Patents as indicators of corporate technological strength. *Research Policy, 16*, 143–155.

Rivitte, K. R., & Kline, D. (2000). *Remnants in the attic: Unlocking the hidden value of patents*. Harvard Business School Press.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of American Society for Information Science, 24*, 265–269.

Stuart, T. B., & Podoly, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal, 17*, 21–28.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *RAND Journal of Economics, 21*(1), 172–187.

USPTO. (2001). Manual of patent examining procedure (8th ed.), August 2001.

White, H. D., & Griffith, B. C. (1981). Author co-citation: a literature measure of intellectual structure. *Journal of American Society for Information Science, 32*, 163–171.

Yulan, H., & Cheung, H. (2002). Mining a web citation database for author co-citation analysis. *Information Processing and Management, 38*, 491–508.