Manuscript submitted to
APDSI 2004 Conference

Generation of technology clusters for patent classification:
patent-citation approach

April 2004

Wook Lee, Byungun Yoon, Yongtae Park*

School of Engineering
Department of Industrial Engineering
Seoul National University
San 56-1, Shillim-Dong, Kwanak-Gu, Seoul 151-742, KOREA

* TEL: 82-2-880-8358, FAX: 82-2-889-8560, E-Mail: parkyt@cybernet.snu.ac.kr

It is confirmed that this item has not been published nor is currently being submitted elsewhere
Generation of technology clusters for patent classification:
patent-citation approach

Wook Lee, Byungun Yoon, Yongtae Park*

Department of Industrial Engineering, Seoul National University, San 56-1,
Shillim-dong, Gwanak-Gu, Seoul 151-742, South Korea

Abstract

The conventional scheme for patent classification is due to the class-based one. Although useful and meaningful to some extent, the class-based scheme is insufficient to meet the needs of specific industries or technological purposes. In response, the primary objective of this paper is to propose an approach for generating technology clusters and suggest a new patent classification scheme. The proposed approach attempts to group patents based on the patent citation information. The main analysis is composed of three parts. First, technology clusters are generated by class-wise patent citation relationship. Second, network analysis is performed to find the relationship among clusters and provide a visual display. Finally, quantitative analysis is executed based on various patent indicators.

Keywords: USPC system; USPTO class; Technology cluster; Patent citation

1. Introduction

With the coming of the knowledge-based economy, the acquisition and development of technological knowledge become more emphasized as the core-capacity for the survival of global competition (OECD, 2001). Hence, various indicators are being developed and used for measuring the capacity and level of technological knowledge (Denison, 1985; Park et al, 2003). Among others, the information in patent data becomes especially important as an indicator of technological knowledge management (Scherer, 1984; Tijssen, 2001; Hu & Jaffe, 2001). As a quantitative indicator representative of the rights and level of specific technology (Pilkington et al, 2002; Ernst, 2003), patent data is piled into databases at national, regional and corporate levels

* Corresponding author
to provide easy access to users. In particular, the United States patent and trademark office (USPTO) database has been widely used as an ample source of technological knowledge because the USPTO database covers all countries and has a relatively accessible class-based structure.

However, USPTO’s patent management is subject to some serious drawbacks. First, the United States patent classification (USPC) system is too general to satisfy the needs for technological forecasting, research planning and technological positioning (Archibugi, 1996). Second, the system is insufficient in reflecting a technological niche (Lai, 2003). Third, the USPTO database tends to generate inexplicit classifications for technologies due to a cumulative characteristic and it is hard to reflect the trend of patent applications and the substantial knowledge flows (Kazenske, 2003). Thus, the classification of new technology is getting harder due to the complex social and economic requirements for new technology, the emergence of fusion technology, and short technology lifespan. Finally, some difficulties exist in looking over related technological knowledge in the current USPC system.

Recognizing the drawbacks of the USPC system, the main objective of current research is to propose a new approach to using the USPC system more efficiently in terms of patent utilization and with a grasp of the current state of technology. The new approach is based on the grouping of the USPTO class through the patent citation information at the class level. As patent citations, the count of citations of a patent in subsequent patent, can be used as an indicator of knowledge flow (Jaffe, 1999), it applies to this research to measure substantial technological knowledge flows. The main analysis can be divided into three parts. First, we generate technology clusters based on the relative frequency of patent citation at the class level. Second, we conduct a network analysis for identifying the relationship of among clusters and providing a visual network of clusters. Finally, we analyze each cluster quantitatively through various patent indicators. By this approach, this research will show the knowledge flow among technology clusters visually and quantitatively. Furthermore, such results can provide useful insights to using the USPTO database more efficiently.

This paper is organized as follows. First, the theoretical background of this research is presented in section 2. In section 3, the overall process of this research is described. Next, the process of generating technology clusters based on patent citations and conducting a network analysis and patent indicator analysis is described in section 4. Finally, implications of current research and issues of further research are discussed.
2. Background

2.1 United States Patent Classification (USPC)

The UPSTO database, one of the major patent systems along with the European patent office (EPO), covers all countries and can be used for monitoring technological activity, specialization and internationalization (Tijssen, 2001). The USPC system, USPTO classification system, has evolved for more than one hundred years. The USPC system is based on the technology disclosed in US patents and uses classes, which contain smaller subclasses, as primary search units (Falasco, 2002). Now, the USPTO database contains 432 classes (3-digit) and about 130,000 sub-classes (6-digit). Each class and sub-class represents specific technologies and they are categorized using key rationales based on utility and structure (Falasco, 2002). To place patents in class categories, systematic but empirical methods are used by examiners. Hence, in patent classification, the more complex case, the harder it is to classify (Falasco, 2002). Another USPC system issue is in discord with the International Patent Classification (IPC). Generally, the USPC and the IPC are aware of best-known patent classifications in current use. However, there are some cases where the USPC and IPC are assigned differently to the same kind of technology. As the major benefits of the use of patent classification for information retrieval are providing access to concepts and overcoming language barriers (Adams, 2001), the discord between USPC and IPC can vex users.

In sum, the USPC system is relatively well-organized and provides useful information to users on patent management as one of the major patent systems. Furthermore, the USPC system generally appears to be more responsive to subject matter changes (Adams, 2001). However, this system still has some shortcomings, such as time consuming patent classification, ambiguity of technology classification and the discord between USPC and IPC. Hence, many studies, such as those on automation of the USPC system (Smith, 2002), analysis of the difference between the IPC and the USPC systems (Adams, 2001) and the reforming of IPC (Calvert, 2001; Makarov, 2004), have been conducted to overcome these shortcomings.

2.2 Patent citation

Patent data have been used as a substitute indicator for technological knowledge, since patents cover the majority of technological fields, as well as a large number of product fields and industrial sectors (Tijssen, 2001). Those patent data are accumulated into databases over a long period and enable the creation of various indicators based on the amount of patent data, patent citations and patent claims etc. Using these indicators, patent information can be used for competitor monitoring, technology assessment and
R&D portfolio management (Ernst, 2003). Patent citation analysis becomes especially important in the analysis of technological knowledge flows. Patent citations are generally defined as the count of citations of a patent in subsequent patent or non-patent literature, and it can be used an indicator of technological impact of the patented invention. Hence, a highly cited patent is likely to contain an important technological advance. Based on this idea, patent citation analysis uses bibliometric techniques to analyze various information of the patent citation, searching for the linkage between cited and citing patents (Karki, 1997). Therefore, the patent citation analysis can apply to the identification of cutting-edge technological activity, technological mapping and competitive intelligence and it produces such useful technology indicators as highly cited patents and technology cycle time and so on.

For these reasons, patent citation analysis has been used in the measurement of technological development (Pilkington, 2002), the analysis of science-technology interactions (Tijssen, 2001) and patent-based cartography of technology (Engelsman, 1994). However, patent citation analysis has some shortcomings. First, it merely indicates individual links between two particular patents, which limits the scope of analysis. Second, it may produce superficial or misleading indicators due to taking only existence or frequency of citations into account (Yoon, 2004). Nevertheless, patent citation analysis has still been used broadly as an analytical tool for technological quality and influence.

3. Methodology

3.1 Basic concept

The main objective of current research is to generate technology clusters for using the USPC system more efficiently. For this purpose, the new approach uses the class-wise patent citation information. As patent citations can be generally used as an indicator of technological knowledge flows, we assume that the frequency of cited-patents from a specific class to others represents the degree of knowledge flows among classes. Ultimately, the research concept is that patent citation relationship can be applied to the technological knowledge links between classes which include those patents. And we regard the degree of the knowledge flow based on the patent citation relationships as a measure of similarity amongst clusters. Figure 1 depicts the basic concept of this research.
3.2 Overall framework

To generate and analyze technology clusters, the research goes through several steps. First, data collection and data preprocessing are conducted as the preliminary step. Based on the collected patent citation data from USPTO database between 1975 and 1999, the citation relationship matrix among classes is generated. Second, the cluster analysis is applied in the grouping of USPTO classes. Third, network analysis is performed in order to show a visual display and examine relationships among clusters. Finally, based on some quantitative patent indicators, in-depth analysis is performed to obtain useful insights. Figure 2 shows the overall process of the proposed research.

3.3 Major tasks

3.3.1 Generation of technology clusters

The USPTO classes organized for each technology are based on a practical analysis of that technology contained in the US patents. However, it is hard to identify whether those classes are classified at the equivalent level. Hereupon, we group the classes into some clusters using patent citation relationship for measuring the relationship among classes. As patent citation relationship generally illustrates how closely related some technologies are to each other, these clusters can be addressed as technology clusters.
3.3.2 Generation of technology network

To obtain the relationships among technology clusters and its features, network analysis is performed. Generally, network analysis, derived from graph theory, facilitates the analysis of interactions between actors. In this analysis, technology clusters become nodes (actors) and knowledge flows among clusters are considered as links (edges). Through the network analysis, a visual network among clusters is exhibited and analyzed by such quantitative indexes as node centrality and betweenness.

3.3.3 Analysis of technology clusters

Various patent indicators are used to compare technological features of each cluster. As we basically use patent data to generate clusters, the features of clusters should be explained in the technological dimension, not economical or social. Thus, technological features of clusters are analyzed by such indicators as highly cited patent, technology cycle time and assignee.

4. Results of application and analysis
   
4.1 Data

The data source of current research is patent citation data from the National Bureau of Economic Research (NBER). The database comprises detailed information on the U.S. patents granted between 1975 and 1999. From this, we collect almost 13,965,000 pairwise citation data, including 3-digit class number and detailed information.

4.2 Cluster analysis

To do cluster analysis, we collect the frequency of class-wise patent citations and generate the citation relationship matrix (432 by 432) among classes. Then we modify the initial matrix by removal of the self-citation frequency because most classes cite their own patents in a large percentage. By applying the citation relationship matrix as input, cluster analysis is performed. The current research applies the hierarchical classification clustering algorithms. Among various possible numbers of clusters, the ratio of intercluster distance to intracluster distance as the measure of the best clustering outcome is obtained. Here, the intercluster distance is defined as the average distance between cluster centers and the intracluster distance is defined as the average distance between cluster center and each element. Hence, the best clustering outcome tends to show the high value of this ratio. Table 1 shows the comparison of this ratio among some selective numbers of clusters.
According to comparison, the case of 30 is selected to generate the best clustering outcome. Consequently, all classes in USPTO are assigned to 30 clusters. The clustering result of some selective clusters is shown in Table 2.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Affiliated classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Apparel, Foundation garments, Surgery</td>
</tr>
<tr>
<td>C2</td>
<td>Power plants, Refrigeration, Fluid handling, Heat exchange, Motors, Heating systems, Pumps</td>
</tr>
<tr>
<td>C3</td>
<td>Beds, Tent, Carriers, Receptacle, Jewelry, Miscellaneous hardware</td>
</tr>
<tr>
<td>C4</td>
<td>Compound tools, Ordnance, Amusement devices, Bookbinding, Semiconductor</td>
</tr>
<tr>
<td>C5</td>
<td>Drug, food, Metal treatment, Tobacco, Sewing, Electric heating, Chemical</td>
</tr>
<tr>
<td>C6</td>
<td>Boot and shoe making, Whips, Brush and broom</td>
</tr>
<tr>
<td>C7</td>
<td>Textiles, Plastic and nonmetallic article, Coating, Fabric, Synthetic resins or natural rubber</td>
</tr>
<tr>
<td>C8</td>
<td>Distillation, Gas, Chemistry</td>
</tr>
<tr>
<td>C9</td>
<td>Metal deforming, Tools, Woodworking, Cutting, Tool changing</td>
</tr>
<tr>
<td>C10</td>
<td>Data processing, Television, Optical, Signals and indicators, Communication, Electricity</td>
</tr>
<tr>
<td>C28</td>
<td>Organic compounds</td>
</tr>
<tr>
<td>C29</td>
<td>Synthetic resins / natural rubbers</td>
</tr>
<tr>
<td>C30</td>
<td>Synthetic resins / natural rubbers</td>
</tr>
</tbody>
</table>

In assumption of analysis, the affiliated classes within each cluster are considered as similar elements in terms of technological knowledge flows. Based on this assumption, some qualitative features of technology clusters are reviewed. First, there are two types of technology clusters. One is a heterogeneous type which includes various classes, and the other is a homogeneous type which has one or several classes. C2, C3, C5, C7 can be categorized into the former, while C28, C29, C30 into the latter. This feature, especially in the heterogeneous type, can be used to explore the possibility of fusion technology. Second, even same technologies can be classified into different clusters. In the case of ‘Synthetic resins / natural rubbers’ class, they are divided into C29 and C30. This result indicates that hierarchical level of technological classes is not consistent. Hence, this may be the cause of misclassification results in further difficulties for patent
management. Third, the class title is hard to understand the features of class. Since C7 includes mostly ‘Textile/fabric technologies’, the ‘Apparel’ class nominally implies close relationship with textile-related technologies. However, the above results explicitly suggest that the ‘Apparel’ class is considerably overlapped over ‘Surgery-related technologies’ as C1. This result is due to the gap of patent citation quantities at the subclass level and the obscurity of subclasses. In practice, the ‘Apparel’ class is broadly related to textile fabric technologies as well as ‘Surgery-related technologies’. However, because both ‘Surgery’ and ‘Apparel’ classes deal with ‘Body protecting/Covering technologies’ in a large percentage as subclasses, the two classes seem to have similar knowledge flows. Thus, class structure may need to be adjusted and class title also needs to be more specific. Finally, technology clusters can be interpreted logically in the viewpoint of industry or product. The cluster, C8 nominally implies “Refinery industry technologies’ and C10 tend to be ‘Image processing-related technology’ in general. Thus, we may name the technology clusters reasonably.

4.3 Network analysis

Based on technology clusters, the network analysis is performed to illustrate the relationship among clusters and understand their features. For this purpose, we use UCINET 6, a networking software package for network and graph. To illustrate the most visible and meaningful network, we select the cut-off value of 10% by sensitivity analysis. That is, if the relative quantity of knowledge flows from specific class to others is greater than 10%, we consider it as considerable knowledge flow. The network is shown in Figure 3.

In the Figure 3, each node represents clusters and the links indicate relation among clusters. Also, the arrow shows the direction of knowledge flows from cited clusters to citing clusters. In addition, the closer distance amongst clusters, the more related they are. To explain the overall network structure and the features of each cluster within network, we perform the quantitative analysis of cluster network using the network centrality index, developed to measure the relative centrality of each node in viewpoint of whole network. First, the relative importance of cluster is measured by gauging the node centrality which represents the density of linkage with other clusters. In this research, we define the node centrality index \( C_D \) as:

\[
C_D(n_i) = \frac{d(n_i)}{g - 1}
\]

\( d(n_i) \): Number of links of cluster \( i \), \( g \): Total number of clusters

Hence, high node centrality value shows the relatively strong influence on other clusters.
Second, the betweenness centrality is used to measure the intercluster influence on the others. Hence, if a cluster has a large betweenness centrality, implying that to have some control over path in the network, the cluster must be between many of the clusters via their geodesics. Using these indexes, we can categorize clusters in terms of the features of technological knowledge flows. These results of some selective clusters are represented in Table 3.

<table>
<thead>
<tr>
<th>Indexes</th>
<th>$C_D$ (Outflow)</th>
<th>$C_D$ (Inflow)</th>
<th>Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>High value</td>
<td>0.79 (C5)</td>
<td>0.14 (C10)</td>
<td>0.19 (C10)</td>
</tr>
<tr>
<td></td>
<td>0.37 (C10)</td>
<td>0.14 (C11)</td>
<td>0.16 (C7)</td>
</tr>
<tr>
<td>Low value</td>
<td>0.00 (C6)</td>
<td>0.03 (C5)</td>
<td>0.00 (C6)</td>
</tr>
<tr>
<td></td>
<td>0.00 (C29)</td>
<td>0.03 (C29)</td>
<td>0.00 (C29)</td>
</tr>
</tbody>
</table>

These result show that cluster C10 has the highest values of each index, while cluster C29 shows the lowest value. As represented in Table 2, cluster C10 includes more varying and fundamental technologies, such as chemistry, metallurgical and electricity-related technology, than other clusters, so the high values of C10 can be explained. However, like C29, the clusters which include only one or a few classes tend to have low values, implying that they have weak influence on other clusters. In case of C5, it has high values of $C_D$ in outflow but low values in inflow. This result is interpreted that some clusters act as the knowledge supplier having strong outflows. And, in the
opposite case, clusters such as C29 having only inflows, can be the *knowledge user*. Furthermore, using the betweenness index, we can find that some specific clusters have some controls of knowledge transmission over the network. Thus, these clusters can be addressed as the *knowledge intermediary*. Based on this idea, we can categorize clusters into the *supplier*, *user* and *intermediary* of knowledge. In Figure 3, white circles represent the *knowledge supplier* while black circles show the *knowledge uses*. Also, gray indicates the *knowledge intermediary*. The notable result is that C10, ‘Image and data processing technology’, has all kinds of characteristics. This indicates that those technologies play a major role of creating other technologies.

### 4.4 Patent indicator analysis

#### 4.4.1 Highly cited clusters

In general, highly cited patents tend to be more important than others in the viewpoint of technological impact (Karki, 1997). Based on this idea, highly cited clusters can be more influential on technological knowledge flows. Hence, although the quantities and scope of patents affiliated in each cluster is considerably different, this information can be useful for searching technologies and measuring the tendency of technological knowledge flows. Table 4 shows the citation portion of some selective clusters.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C5</th>
<th>C10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation portion</td>
<td>5.85</td>
<td>9.67</td>
<td>7.33</td>
<td>17.2</td>
<td>14.37</td>
</tr>
</tbody>
</table>

As a result, five higher ranks occupy almost 50% of all citation data. As anticipated, these clusters mostly include basic technologies related fundamental industries. Thus, these results considerably match with other results. Furthermore, it can be used to time-series analysis to measure technology trend.

#### 4.4.2 Technology cycle time (TCT) index

To measure the speed of technical advance or the newness of subject patents, the TCT index is usually used in citation analysis. Technology cycle time is the median age, in years, of earlier patents cited in the new patents (Karki, 1997). Generally, the shorter cycle time, the quicker the changes. Hence, TCT index is applied when measuring the age gaps between the cited clusters and the citing clusters. For this purpose, we measure the average of technology cycle times between cited and citing patents in each cluster as the cluster-wise TCT index. Table 5 shows the result of some selective clusters.
As anticipated, high-technologies such as ‘Image processing technologies (C10)’, ‘Optical technologies (C21)’, and ‘Data processing technologies (C25)’ tend to exhibit short technology cycle time while such conventional technologies as ‘Package technologies (C12)’ and ‘Button-related technologies (C14)’ show the longer cycle times. This phenomenon indicates that the TCT index enables comparative analyses amongst technology clusters, as well as patents.

### 4.4.3 Assignee index

Assignee, which represents the ownership of technologies, can be used as an indicator of the source of technological knowledge in patent analyses. The USPTO classifies patents by type of assignees, into the seven categories. In this research, we reclassify those into four categories, ‘Unassigned’, ‘Corporations’, ‘Individuals’ and ‘Governments’. To confirm technology possession, we examine the proportion of the type of assignees. In this analysis, we use a Chi-Square test for comparing the differences. The results of some selective clusters are shown in Table 6.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Unassigned</th>
<th>Corporations</th>
<th>Individuals</th>
<th>Governments</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3</td>
<td>47.8</td>
<td>49.2</td>
<td>2.3</td>
<td>0.6</td>
</tr>
<tr>
<td>C17</td>
<td>21.4</td>
<td>71.4</td>
<td>4.8</td>
<td>2.4</td>
</tr>
<tr>
<td>C19</td>
<td>40.0</td>
<td>46.7</td>
<td>0</td>
<td>13.3</td>
</tr>
<tr>
<td>C24</td>
<td>6.7</td>
<td>89.9</td>
<td>0.5</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Table 6 indicates that the type of assignees in each cluster is significantly different (α < 0.001). From Table 6, C3, mostly ‘Commodity-related technologies’, shows the highest ratio of the ‘Unassigned’ type while C23, ‘Communication-related technologies’ shows very high ratio of the ‘Corporations’ type. This phenomenon indicates that corporations are eager to have technologies which possess higher commercial values. Also, C19, regarded as ‘Railway delivery-related technology’, has the highest value among the ‘Governments’ type, since the railway-related industry is too large and wide to be handled by individuals and small groups. Thus, the governments own considerable degree in this field. Finally, some specific technologies such as ‘Type casting
technologies (C17)’ tend to show relatively high value of ‘Individuals’ type.

5. Discussion and Conclusions

The operational and strategic utility of USPTO database becomes more apparent. However, the USPC system is insufficient in reflecting technological knowledge flows and its hierarchical structure for technology classification is not consistent. In order to overcome the flaws of the USPC system and assess the USPC system in the technological knowledge flow aspect, this research applies cluster analysis and network analysis based on patent citations to the USPTO class.

The research is largely composed of three parts, generating technology clusters, visualizing a cluster network and performing a quantitative analysis. As a result, 30 clusters are generated and the features of each cluster are analyzed through network analysis and patent indicator analysis. The results of this approach provide some meaningful information about the USPC system. First, recognizable groups of classes, namely technology clusters, can be generated in viewpoint of substantial knowledge flows. Hence, it suggests the direction of the improvement for USPC system. Second, through the visualized cluster network, the overall relations among technology clusters are understood easily. This helps users of USPC system to understand relative state and importance of technologies. Finally, the result shows influential technologies and the linkages of heterogeneous technologies. Hence, it assists users in finding the possible areas of fusion technology.

In the analytical process of this research, however, the proposed method is still subject to some limitations. First, in some cases, the naming of clusters may be difficult due to the variety of technologies. Second, patent citations, a basis of this research, still does not reflect the most recent technology trends due to the time-lag of citation. Nevertheless, this approach is still meaningful for the management of patent and technological knowledge flows. With further research, this approach can be extended far beyond the scope of current research. For example, proposed methods can be applied to the more detailed technologies, the USPTO subclass, in specific technological field (Class). In addition, dynamic analysis is being considered for measuring the periodic technology clusters and technological trend over time.

Acknowledgement: this research was funded by the National Research Lab (NRL) program of the Ministry of Science and Technology of Korea.
References
