Generation and Application of Patent Claim Map: Text Mining and Network Analysis
Juneseuk Shin and Yongtae Park†
Department of Industrial Engineering, Seoul National University, San 56-1, Shilim-Dong, Kwanak-Gu, Seoul, 151-742, Korea
Received 3 November 2004

Despite the fact that patents are under intensive scrutiny for years, patent claim, the most ample source of information has been relatively unexplored. Patent claims mean the right over a patent. Their overlaps by subsequently granted patents indicate the erosion of patent rights. In that regard, the issue of patent valuation and competitor strategy is very closely related with it. In addition, claims could be used to recognize technology relatedness. Therefore, in this research, an exploratory method to deal with patent claims using text-mining and network analysis has been proposed. First, a claim overlap profile is constructed to identify whether a specific claim overlaps another by applying text mining and domain expert knowledge. Secondly, network analysis is used to generate three kinds of patent claim map. This could help researchers, R&D managers and policy makers to evaluate patents and analyse competitors more accurately, and develop patent strategy more efficiently. In the long run, the patent claim profile and map could contribute to the overall technology management including new technology development, strategic positioning of technology and technology alliance.

Keywords: Patent claim, valuation, competitor, network analysis, text mining

Technological change has had a decisive impact on the competitive structure of many new industries. In fact, it has become key factor of competitiveness among companies, and as a consequence, companies are now investing enormously on R&D. Thus, with the increase in the importance of technology management, technology assets such as product/process data and patents have become the centre of attention. Especially, patents and patent statistics have been under intensive scrutiny by researchers because of their important role in technology management. Furthermore, they facilitate an analytical work due to their relative advantages with regard to availability of database, scope of coverage and richness of information.

Narrowing focus on a corporate setting, patent analysis has been used mainly to evaluate the competitiveness of a firm, analyse competitors’ strategic moves and explore opportunities to develop new technologies. Yet, all these depend on two prerequisites, the value of a patent and relatedness among patents. The former has attracted researchers’ interests for years because it enables strategic maximization of a patent portfolio, facilitates transaction of patent, etc. Theoretical researchers have discussed value-determining parameters of patents for a long time. After Nordhaus first reported on the lifetime of a patent, similar researches have ensued. Klemperer and Gilbert and Sapiro introduced the patent breadth as a new parameter. Greene and Scotchmer elaborated the concept of novelty and inventive activity, entertained by Nordhaus. On the other hand, industrial researchers have persevered to develop substantial value indicators using renewal and citation approach. The latter is twofold, backward citations and forward citations.

In addition to them, various factors have been explored such as scope, opposition annulment and family size. These are all proxy measures that partially reflect technology information of a patent. Nevertheless, naïve use of patent statistics is still continuing due to the fact that there are no practical alternatives. Also, even though claims provide more ample information than renewal, citation, etc, researchers have faced many difficulties in handling patents properly.

Patent relatedness also plays a key role in recognizing a firm’s strategic position and helps to assess the competitive potential of a certain technology. There have been numerous attempts to find appropriate measures. First, Jaffe proposed a
method of using the distribution of patents over 49 technology fields classified into 12-digit IPC (International Patent Classification) code. Then, recently, some bibliometricians refined this approach to include measurement of co-occurrence of classification codes including not only a primary code but also a supplementary code. There were many others who adopted citation analysis to link patents. This is very similar to that used in science and involves linking references in a scientific paper database. Citation has been used as a proxy measure of knowledge flows or technology linkages. Although these are useful methods to arrange and visualize raw patent data, they do not guarantee full use of patent data. In fact, it should be noted that a patent is linked to a certain technology class according to the claims and not through citation relationship. This entails great loss of information if patent claims are overlooked. Classification code is at best information compressed from original documents. There are also some drawbacks to citation analysis. Above all, the scope of analysis and the richness of potential information are limited because citation analysis takes only citing–cited information into account.

In this paper, an exploratory method to generate patent claim maps by employing text mining and network analysis is used. First, the number of claims of a patent overlapped with other patents is determined using text mining and domain expert knowledge. Here, citation and similarity of technology classes play a supplementary role. Using the results, three kinds of patent claim map can be generated. The information and insights extracted from it are crucial to a firm’s patent strategy, more broadly to technology management.

Methodology

Text mining is used as the screening method. Network analysis combined with domain expert knowledge is the main tool to extract, arrange and visualize useful information from these overlapping claims among patents.

Network Analysis

The interactive relationship among a set of actors could be expressed as a network. Network analysis basically aims to analyse directed or undirected flows (edges) between actors (nodes). The structure of relationships among actors and the location of individual actors in the network provide rich information on the behavioral, perceptual, and attitudinal aspects of individual units and the system as a whole. The applicability of network analysis is very wide. Some typical topics include the inter-industrial diffusion and adoption of innovations and the human network in knowledge management.

In the context of patent analysis, individual patents are allocated to nodes and their relationship among patents is accounted for edges. Relationship could be attributed to a variety of aforementioned variables such as citation, co-occurrence of classification code, etc. The patterns in a linkage and locations considerably vary according to choice of a relationship variable. Moreover, the overall landscape could be changed if looked from a different perspective by applying other graphical algorithms that give new information and insight about the relationships among patents.

Text Mining

Data mining applies to a machine-learning algorithm for extracting information and its patterns in a database. The level of the efficiency and usefulness of data mining depends on the structure of the database. This technique has limitations in its utility in the case of huge unstructured documents. Text mining is a newly developed technique that makes use of keyword vectors extracted from documents. Basically, it puts a set of labels on each document on which discovery operations are performed. By doing so, documents could be featured by keywords. Recently, text mining has attracted increasing interest and has been actively applied in knowledge management.

Text mining is particularly useful in patent analysis because patent documents are typically roughly structured. It is inevitable to transform raw data into structured one. Previously developed patent statistics are valuable, but insufficient. Thus, measures such as a keyword vector or frequency using text mining will help in approaching patent from a different angle. But, in order to minimize the loss of information in the process of extracting or compressing information, it is necessary to develop an advanced text mining technique.

Data

Useless results are obtained when text mining is applied to huge documents. To begin with, care must
be exercised in choosing the period of data because the overlaps of claims are most likely to be fewer in sparsely distributed ones. At the same time, it is important to control the size of data properly. If the size is too big, text mining will not produce useful screening results. In addition, the characteristics of a technology such as its lifecycle, etc, should be taken into consideration. For instance, for old technologies such as agriculture-related ones overlaps of claims are unlikely. However, in emerging technologies, there are very few data to analyse. Considering the foregoing facts, the technology of data processing that covers financial, business practice, management or cost/price determination must be selected. This technology conforms to US Patent and Trademark Office (USPTO) patent class 705. The period from 1991 to 2001 is selected as the number of patents suddenly increased afterward. The quality and content of technology also showed a shift in technology trends. So, these have to be analysed separately. In this research, only the former period is dealt with. Raw patent documents are extracted from USPTO database.

**Research Framework**

The overall process of generating patent claim map is shown in Figure 1. Raw patent documents in electronic text provided by USPTO follow a natural language format. Moreover, claims are mixed with other information such as description, assignee, etc. Thus, claims were manually separated and unnecessary information was eliminated. Then, by applying text mining to this set of patent claims, keywords were extracted from the database and each patent document was expressed as a corresponding keyword vector. Based on its similarity, patents were classified into several clusters. Picking up a certain patent of interest, overlaps in the claims with other patents were thoroughly investigated. Firstly, domain experts counted the number of overlaps within the cluster to which the patent belonged. Secondly, the adjacent or similar classes were also examined. Finally, by tracing the citation relationship, some overlaps could be found. Combining all these findings, a claim overlap profile that demonstrates the number of claims overlapped across several technology classes could be constructed. Delving into patent-to-patent relationship, three kinds of patent claim map were generated. The first map simply visualized a claim overlap profile and added information of assignees. In the second map, all the relationships of patents on the network were added. Lastly, by emphasizing a group of nodes that displayed great intensity of ties, the third patent claim map was generated using K-core. Some information in more detail within same cluster was arranged to help more precise competitor analysis.

**Claim Overlap Profile**

Even within a certain patent class, it is almost impossible to verify all the overlaps among patents. Thus, it is only feasible to build a set of patents that are most likely to have overlapped claims of another patent. Text mining is a good technique to screen unrelated or least related patents at the first stage. Using the Textanalyst 2.1, 627 keywords were extracted from 252 patents. During this process, extra care was taken with respect to the frequently or rarely appeared ones which are of little use as a classifier. Applying the cutoff combining frequency with weight, 205 keywords were obtained. Next, the keyword vector was constructed. If a specific keyword is included in the claims of a patent, the corresponding keyword vector column is filled with the frequency in which that keyword appeared.

In connection to keyword vectors, it was necessary to narrow down the patents to analyse using clustering analysis method. In applying this method, there are some conditions that need to be considered before fixing the cluster. Firstly, the intra-cluster similarity of patents must be maximized while ensuring that the inter-cluster difference remains within a proper level. Secondly, the size of cluster should neither be too big nor too small, otherwise the cluster becomes entirely useless. Considering these factors, it was decided to
adopt K-means clustering analysis. Among the various association indexes, the common Euclidean index was used in this research\textsuperscript{35}.

From the clustering analysis, 17 clusters were obtained (Table 1). It can easily be seen that some of the clusters are meaningless because the number of patents is too small. Nine of the clusters can be eliminated by applying the cutoff value of nine. The remaining eight clusters were used for examining overlaps among claims. This first screening stage is meaningful in that it saves the trouble of dealing with unstructured patent documents.

At the second stage, domain expert knowledge plays a key role in judging the similarity of claims. The problem is how to calculate the similarity of claims between patents when the level of importance of claims is not uniform. A certain patent may have just two or three claims that are very comprehensive. On the other hand, some patents have 30-40 principal and trivial claims. Sentence mining, which could calculate the one-to-one similarity between sentences, still does not help in this kind of detailed comparison. And neither is the keyword vector mining any better than it. So, the number of similar claims was counted and divided into two categories, ‘similar’ and ‘partly related’. The patent bearing number 5, 732, 397 and granted in 1992 was chosen as an exemplary case from the 15\textsuperscript{th} cluster through examination. The claim overlap profile of this patent in a specific cluster was arranged (Table 2).

There are doubts as to whether the above results are complete sets that encompass all the possible overlaps. Certainly, it cannot be confirmed. USPTO classification system has a superiority rule when claims could be classified in different classes.\textsuperscript{36} In other words, adjacent or similar classes of original classes may share similar patents. Secondly, in many cases, a technology is composed of combinations of numerous technologies. In such a case, a technology class that seems to have nothing to do with the concerned class could have similar patents. This could be identified by using two kinds of information, the most useful one is to test citation relationship among the patents. Patent examiner confirms that a certain patent is closely related to the cited ones that are most likely to have similar claims. The other one is the similarity of technology classes referring to class definition. In order to make up for the profile that was previously missed out, a finalized claim overlap profile of a patent was constructed after adding this information (Table 3). The table arranges the overlaps of claims between the patent 5,732,397 and other related patents.

Six partly related patents could be found in the original class 705. These are the lost ones during the clustering process. Yet, there are no patents that could be classified under ‘similar’ categories. Referring to class definition, it can be concluded that the adjacent class, 706, was worthy to be explored. The entire set of patents was studied, and 3 similar patents and 6 partly related ones were found. Finally, by using the citation relationship, 3 patents that had overlapped claims were added.

**Patent Claim Overlap Map**

With the results on the overlaps among patents, a claim overlap matrix was constructed as a prerequisite for generating a network. The relationship was

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patents</td>
<td>50</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>8</td>
<td>40</td>
<td>23</td>
<td>13</td>
<td>21</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>68</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Patent number</th>
<th>Number of overlapped claims with an original patent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar patents</td>
<td>5, 717, 865</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5, 742, 776</td>
<td>7</td>
</tr>
<tr>
<td>Partly related patents</td>
<td>5, 887, 154</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5, 943, 650</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Class</th>
<th>Similar</th>
<th>Partly-related</th>
<th>Number of claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text mining</td>
<td>705</td>
<td>2</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Class definition</td>
<td>705</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>706</td>
<td>3</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>382</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Citation</td>
<td>434</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>702</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

---

**Table 2 — Claim overlap profile for a specific cluster**

**Table 3 — Claim overlap profile of a patent 5, 732, 397**
quantified in terms of the number of overlapped claims. The erosion of the patent right was in proportion to the number of overlapped claims in this matrix. Using the data in Table 3, an exemplary claim overlap matrix whose size is 23 x 23 was constructed. By applying this matrix as an input, some network software packages can generate networks. In this research, UCINET 6 and Netdraw were used.

A well-constructed network often conveys an intuitive knowledge on the structure of a system. Likewise, the maps were arranged according to the research objective. One of the research objectives is to grasp the overall view of competitors around a specific patent. Blanketing or blocking is a typical patent strategy discussed so far. Using the assignee information of patent documents, a map that gives an overall view of competitors as well as the position of a specific patent could be constructed (Figure 2). Gray and dark gray nodes denote the patents by same assignees, Teknowedge and Merrill Lynch. They have more than two patents in this map. The distance from the centre is determined by the number of overlapped claims among patents. Consequently, this map visualizes conceptually discussed strategic positions among patents.

However, there remains a problem. If a firm wants to buy some patents or make strategic alliance with other firms, it should know the relationship among patents with other neighbouring patents. Figure 2 is generated just on the basis of relation with a specific patent 5,732,397. A prerequisite is that the neighbouring patents should also be evaluated to get insights into the holistic nature of the network. In other words, there is a need to introduce the number of overlapped claims among all patents in the claim overlap matrix (Figure 3).

In Figure 3, the first 3-digit of label at each node indicates the technology class and the rest number indicates the order of data. Same level of shade denotes same cluster based on the similarity of patent claims. One thing that should be noted is that distance is meaningless here. Still, it gives researchers information such as the centrality, coreness, number of linkages and similarity among patents. For example, the patent 7066 shows many overlaps with other patents. This suggests that it is very possible that it will erode other patent claims. From this, the firm can deduce that such a patent is a strong obstacle for a firm to develop related technology. If the distance among patents is stressed as a point and the clusters are classified according to the hierarchical importance, finally a map is generated (Figure 3).

Applying K-core technique with MDS (Multi Dimensional Scaling) method yielded Figure 4. K-cores are not necessarily cohesive subsets but they do identify areas of the graph, which contain clique-like...
In other words, it is a kind of hierarchical clustering based on the information of the number of times each pair of patents are in the same clique. Varying the cutoff value of $K$, stabilized clusters could be obtained at $K=4$. The different shade expressed in white, gray, dark gray and black is used to indicate different clusters by $K$-core with MDS technique. In other words, it suggests that the relationship within a cluster should be examined in more detail. For example, patents of white circles are very important to each other for their owners to develop a patent strategy. MDS makes the distance between patents meaningful. The more overlapped the patents are, the shorter is the distance between them. For delving into the map, it is necessary to construct the cluster profile of white nodes to identify more substantial relationship (Table 4).

Through this profile, key players related to a specific technology can be identified. For instance, principal technologies in this cluster are hierarchical knowledge database and rule-based decision process. Teknowledge is a key player with regard to developing knowledge database. On the other hand, Westinghouse Electric Corp has the technology of rule-based automation. Combining the above information, a manager of a firm can increase the quality of decision concerning new technology development, patent application and opportunities of strategic alliance.

**Conclusion**

Recently, patents have attracted considerable attention of researchers, policy makers and R&D managers. The potential of patent data has been highlighted more as the scale of economy became larger. While the process of technology development became more complex, its lifecycle became much shorter. Various methods have been developed to make use of information within patent documents. Still, the proxy measures such as citations, renewal fees and classification codes are not enough to maximize the utility of patents. Even though patent claims are the most ample source of information, they are relatively unexplored because there is no appropriate method to handle them properly. In this research, an exploratory process of extracting information from patent claims through text mining and network analysis is proposed.

<table>
<thead>
<tr>
<th>Number</th>
<th>Assignee</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>7056</td>
<td>Individual</td>
<td>Method of determining the premium for writing a policy insuring</td>
</tr>
<tr>
<td>7062</td>
<td>Teknowledge</td>
<td>Basic expert system tool</td>
</tr>
<tr>
<td>7065</td>
<td>Teknowledge</td>
<td>Hierarchical knowledge system</td>
</tr>
<tr>
<td>7066</td>
<td>Westinghouse Electric Corp</td>
<td>Automated rule based process control method</td>
</tr>
</tbody>
</table>
By using text mining based clustering and domain expert knowledge, a claim overlap profile of a specific patent can be generated. Citation relationship and similarity of technology class compensate for what the former analysis has missed. By adding the overlaps found in the latter stage, the claim overlap profile can be generated. This profile shows the number claims of a specific patent that are still effective. It also shows the patents that have eroded this patent. Transferring the profile into a matrix form, a claim overlap matrix can be constructed. Based on this, three kinds of patent claim maps can be generated. First map, the assignee map, is developed to provide the overall strategic positions of competitors. In the second one, the overlaps among all patents to evaluate the patents of competitors more accurately and to measure the network characteristics are visualized. The last map and profile help to analyse the technology relatedness and competitors in depth. To sum up, this research contributes to the evaluation of patents, technology relatedness and competitor analysis. This, in turn, contributes to the development of patent strategy of a firm. Furthermore, they can be used to enhance the efficiency of technology management, particularly, in new technology development, strategic positioning of technology, technology alliance and so on.

Although the proposed method is very useful, it still has some limitations. The primary one is how to estimate the importance of each claim and to decide on which claim should be given more weight. If it is possible to decide which claim is more important in numbers, this would substantially help to calculate the extent of overlap. In the aspect of methodology, the accuracy of text mining is also a problem. If the sentence similarity could be calculated accurately, it will also considerably help the valuation. Secondly, claim overlaps tend to pile up over time. Therefore, it has become imminent to develop a dynamic analysis considering the horizon of time. The distribution of overlaps over time is very important with regard to the erosion rate of a patent right. Finally, some advanced indicators with which the above-mentioned information can be compressed would be desirable.

Acknowledgement
This research was funded by the National Research Lab (NRL) programme of the Ministry of Science and Technology of Korea.

References
4. OECD. The measurement of scientific and technological activities: using patents as science and technology indicators—Patent manual (Paris) 1994
7. Archibugi D and Pianta M, Measuring technological change through patents and innovation surveys, Technovation, 16 (9) 1996, 451-468
10. Klemperer P, How broad should the scope of patent protection be? The RAND Journal of Economics, 21(1) 1990, 113-130


Seideman S, Network structures and minimum degree, *Social Networks*, 5, 1983, 269-287