Technological Relatedness based on Co-classification Network Analysis: A Case Study on Electricity Sector

WEN Fang-Fang
School of Management
Henan University of Science and Technology
Luoyang, 471023
China
wenfangfang2013@163.com

JDIM
Journal of Digital Information Management

ABSTRACT*: This paper describes a quantitative study on interdisciplinary technological relatedness represented by patent co-classification. Taking patents in the electricity sector as samples, classification analysis was conducted on the basis of 4-digit codes of International Patent Classification. Then, social network analysis and visualization technique were conducted. The relatedness between different subclasses was measured and visualized. The focus and trend were detected. The results indicate that "H04L: Transmission of digital information" is the most active subclass that can be considered as the focus in the field of electricity. It also suggests that hot domains characterized by high interdisciplinarity may involve some emerging technologies, driven by knowledge relatedness and characterized by technological diversification. The results obtained in this study are expected to provide an improved understanding of technological development and evolving characteristics, as well as provide insights that can contribute to further research and development.*

Subject Categories and Descriptors
H.3.4 [Systems and Software]; Information networks: **D.2.12 [Interoperability];** Data mapping

General Terms: Patent, classification, co-occurence; Information analysis, SNA

Keywords: Co-classification analysis, Technological relatedness, Social Network Analysis, Visualization

Received: 31 August 2015, Revised 28 September 2015, Accepted 5 October 2015

1. Introduction

Intense technological convergence and diversification has resulted in growing relatedness among various technological fields (TFs). This phenomenon represents a megatrend in the advancement of science and technology (S&T) in the 21st century. This condition has also transformed research and development (R&D) from a disciplinary to an interdisciplinary or multidisciplinary pursuit [1]. To cope with this megatrend, governments and firms must investigate the interdisciplinary characteristics of a technology, as well as identify the focus, relatedness, and trends so that new strategies and policies can be formulated and implemented. Since patent documents contain a large volume of technological and commercial information and they significantly influence technological development, patent analysis can be regarded as one of the most effective methods to keep up with technological trends [2].

According to worldwide patent statistics, from 2011 to 2015, China was ranked first in the world in the number of patents applications. Most inventions completed by Chinese firms and individuals are filed at the State Intellectual Property Office (SIPO) of the People's Republic of China, and their documents are primarily included in the local database instead of foreign databases. Therefore, conducting a specialized study on the technological relatedness in China based on local data samples from SIPO is necessary to measure and visualize the technological interdisciplinary situation in the country.

Electricity sector is one of the most active technological fields (TFs) characterized by high interdisciplinarity, with a large number of inventions concentrated during the past decades. Thus, this study took the TF of electricity as the example and the SIPO patents in "*Section H: Electricity*" as the sample. Classification analysis was conducted on the basis of 4-digit codes in the International Patent Classification (IPC, 8th ed.), to measure and visualize the relatedness between various subclasses, as well as to detect the focus and trend.

2. Literature Review

(1) Patent Classification and Co-classification Analysis

A patent classification refers to the manner by which examiners at a patent office arrange patent documents according to the technological features of inventions. Since the same patent may be categorized under two or more classes, the co-classification information can be used to identify the relatedness between different technologies. If an increasing number of patents are categorized under two classes, technologies in these two classes may have more in common and may be becoming closely interrelated. By adopting the degree of co-occurrence of classes, one can quantitatively analyze the interdisciplinarity of technologies [3]. That is, the co-occurrence of classification codes represents an overlapping relationship among various TFs and may also reflect their relatedness, which is the major driver of technological diversification and convergence [4]. Over the past decade, patent co-classification analysis has been widely used to investigate the similarity and relatedness among various TFs, as well as to analyze and visualize the relations among technologies at different levels of aggregation [5]. Moreover, patent co-classification analysis can provide practical information on the technological convergence process and future directions by examining potential knowledge flows among technology classes or subclasses [6].

(2) Social Network Analysis and Visualization

Social network analysis (SNA) outputs a visual as its result, and this visual displays the overall linkage structure among the nodes, which would be difficult to identify using data alone [7]. In addition, SNA allows for identification of local and global patterns, determination of influential nodes, and examination of network dynamics [8]. SNA is pervasive in the social and behavioral sciences, and it can be employed in several other formal sciences, including technology analysis, because of its wide applicability [9].

Mapping technique (also known as visualization) is mainly used to display various aspects of S&T through different indices. The mapping method offers a unique way to visualize developments in TFs and in the overall technological field [10]. The mapping method is often applied with network analysis. Both of these methods are based on relational indicators, which reflect the network of structural features inherent in the scientific or technological activities as represented through scientific studies or patents.

Patent network analysis is an advanced technique for patent analysis, which is based on the co-occurrence of different indicators, including keywords, classification codes, inventors, assignees, and citations [11]. Along with the mapping technique, patent network analysis can visually display the relations between different patents and their indicators. Furthermore, the overall structure and the influential patents can be detected, and the technology tendency can be monitored.

Co-classification analysis combined with mapping technique is a popular method of patent analysis. The existing research has focused on the United States, Japan, and Europe with the sample data obtained from predominantly English-language sources, including Derwent Innovations Index, US Patent and Trademark Office, European Patent Office, and World Intellectual Property Organization. However, few of these sources provide local data samples that concentrate on the domestic situation in China.

With regard to the aforementioned limitation, this study explored the SIPO patents under "Section H: Electricity" as the sample, conducted empirical analysis on patent classification codes and their co-occurrence network to identify the focus and trends, as well as investigated the relatedness between various domains in the TF of electricity in China.

The rest of this paper is organized as follows. Section 3 describes the sample data and methods used in this study. Section 4 presents the research results, including statistical data and visualized maps. Section 5 summarizes the conclusions.

3. Data And Methods

3.1 Data

(1) Patent classification

IPC is the most popular classification system worldwide and is adopted by most countries in analyzing technology development [12]. The IPC system is hierarchically structured and can be used at different hierarchical levels. It has been applied in China since 1985. Currently, it is in its eighth edition (IPC-8), which uses a 12-digit code that includes approximately 70,000 categories.

Patent classification code is assigned carefully by patent examiners of the issuing patent offices. According to the IPC system, a patent may be classified at least by one (main or primary) classification code, but usually more classification codes (secondary or supplementary) are assigned to specify the technological content in detail. The main code refers to the object of claimed and appropriable knowledge and technology, while the supplementary code refers to additional and detailed information to further define the technological content of the patent and the TF to which the patent belongs.

A full IPC code contains 12 digits, which can be divided into 5 levels, namely, section, class, subclass, main group, and subgroup. Section H contains 6 classes labeled with 3 digits and 50 subclasses labeled with 4 digits. Many studies, however, have exploited the 4-digit IPC code, which contains section, class, and subclass, because the 4-digit IPC code shows sufficient technological characteristics related to a patent [13]. This study also chose the 4-digit IPC codes of "Section H: Electricity" as the sample data to conduct empirical research. The symbols and titles of these 4-digit IPC codes are listed in the Appendix.

(2) Retrieval Strategy

The SIPO database is the largest collection of Chinese patent documents, which includes almost all patents issued by SIPO from China and other countries [14]. This study only chose the patent inventions completed by the firms and individuals (subject to the first assignee) from mainland China as the sample data set.

To define the retrieval period, application dates are used instead of issue or publication date because each patent must experience a long period of examination and the application time is much closer to the birth time of a new invention. Our retrieval time is April 2015 with the retrieval time span from 2006 to 2014. Since the classification system currently in use (IPC, 8th edition) was implemented in January 2006. And the patents in 2015 were not included in the database. Patents with application dates in the last two years (2013 and 2014) are not completely covered at the time of retrieval because of the review cycle, as well as the time lag between publication date and the availability in the database.

For "*Section H: Electricity,*" a total of 218,634 patents of inventions, patents of industrial designs and industrial product appearance design excluded, are found for the period from 2006 to 2014, with 75.31% of them having two or more IPC codes. The 218,634 patents contain a total of 540,026 IPC codes composed of 425 different codes. On average, 2.47 IPC codes are assigned for every patent.

3.2 Methods

(1) Frequency Statistic

Firms, institutions, and individuals from all countries seek legal protection for their innovations; a patent is a valid and direct indicator of technological competencies and innovation capacity [15]. The fact that a firm applies for a patent in a given TF means that such a firm is at, or close to, the technological frontier and has advanced its technological competencies in that particular TF. Generally, the more patent codes concentrated on a TF, the more active the TF is. This study examined the frequency statistic on the 4-digit main/primary IPC codes to identify the hot domains.

(2) Co-classification Analysis

As described, each patent is classified at least by one main/primary IPC code and more secondary/supplementary IPC codes. Since one classification code represents one technological domain, the frequency by which two classification codes are jointly assigned to the same patent can be interpreted as a proximity index of the strength of the technological relatedness between different domains. On the basis of such an assumption, a co-classification matrix was constructed for use in the network analysis and mapping.

(3) Mapping Technique

The co-classification matrix was used as the input base to produce a patent network map displaying the co-occurrence relationship among various IPC codes. The map can assist us in intuitively understanding the structure of the entire field and the relatedness among different domains. In this study, the mapping technique was applied along with co-classification network analysis to identify the focus and trends, as well as to measure the interdisciplinary relatedness in the field of electricity. This study used NetDraw program to produce the co-classification network map.

4. Results

4.1 Identifying the Focus

Given a definite period, a high-frequency IPC code means that the domain is much more active and can be seen as the focus of the entire field. This study counted the frequencies of each 4-digit main/primary IPC codes in the sample. Statistical results are presented in Figure 1.

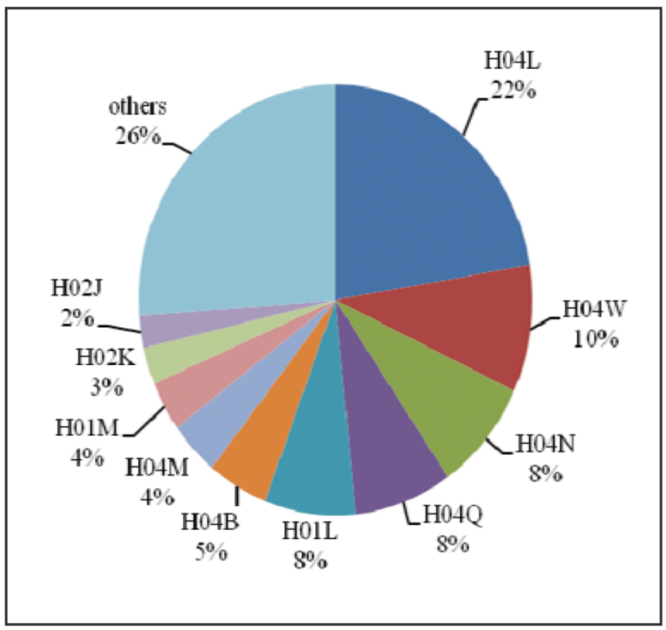

Figure 1. Statistics of IPC codes of subclasses (Top 10)

The proportion can be considered as the indicator that represents the active degree of each subclass. The top 10 frequent subclasses listed in Figure 1 are more active than the others. Among the 50 subclasses in Section H, those 10 subclasses in the Figure 1 above account for nearly 3/4 of patents in Section H, while the other 40 subclasses share the remaining 1/4. The high-frequency IPC codes demonstrate the hot domains in the field of electricity.

Specifically, H04L ranks as the first with a proportion of 22.40% that is far higher than that of any other subclass, indicating that H04L is the most active domain. Some other subclasses, including H04W, H04N, H04Q, and H04L, can be considered as the secondary hot domains in Section H although their frequencies are not as high as H04L.

In addition, among the 10 subclasses in the Figure 1 above, 6 and the top 4 subclasses both come from the class of H04. The top 4 subclasses, namely, H04L, H04W, H04N, and H04Q, account for nearly half of the patents in the entire section. Most patents concentrate on this class; thus, the class of "*H04: Electric communication technique*" is obviously active in the field of electricity. Meanwhile, the subclass of "*H04L: Transmission of digital information*" is definitely the focus of the entire field.

4.2 Measuring the Relatedness

Relatedness between TFs can be measured by co-classification analysis [16]. The relatedness between two TFs (i and j), can be measured by R_{ij}, which ranges from 0 to 1. R_{ij} increases as i and j are co-classified more often. $R_{ij} = 1$ when every time one TF is assigned to a patent, the other TF is also assigned to the same patent. $R_{ij} = 0$ when i and j are not co-classified at all.

$$R_{ij} = \frac{C_{ij}}{\sqrt{N_i * N_j}} \qquad (1)$$

Where C_{ij} is the co-classification frequency of TFi and TF$_j$. N_i is the number of patents classified in TFi while N_j is the number of patents classified in TF_j. According to the hierarchical structure of the IPC system, R_{ij} can be measured from macro and micro levels by the indicators of section and subclass, with the IPC codes labeled with 1 digit and 4 digits, respectively.

4.2.1 Relatedness Analysis at Section Level

The IPC system contains a total of 8 sections. For the patents of Section H, their supplementary IPC codes cover all of the 8 sections, which means that section H has co-classified relations with all of the other 7 sections. between Section H and the other 7 sections were calculated (see Table 1).

R_{ij} values in the Table 1 can be used to detect which section has stronger relatedness with Section H. We found that "*Section H: Electricity*" has the strongest relatedness with "*Section G: Physics*" with the R_{ij} value as high as 0.749, followed by Sections C, B, and F. However, Section H has minimal relatedness with Sections E and D.

Section	R_{ij}
G: Physics	0.749
F: Mechanical engineering; Lighting; Heating; Weapons; Blasting	0.260
E: Fixed construction	0.086
D: Textiles; Paper	0.055
C: Chemistry; Metallurgy	0.395
B: Performing operations; Transportation	0.327
A: Human Necessities	0.108

Table 1. Co-classification between Section H and other sections

4.2.2 Relatedness Analysis at Subclass Level

According to the hierarchical structure of the IPC system, the visualization of co-classification at the subclass level shows the technological relatedness in a more detailed manner than at the section level. In this study, the sample contains 425 4- digit IPC codes; 50 of them belong to Section H while the remaining 375 codes come from the other 7 sections. We measured and visualized the relatedness between different subclasses. To effectively present visuals within the limited page, a threshold ($R_{ij} \geq 0.05$) was introduced to filter the less-strong relations (see Figure 2)

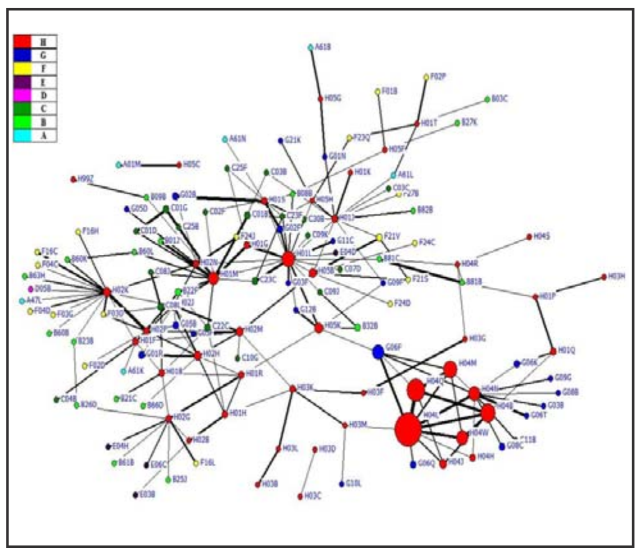

Figure 2. Visualization of co-classification among various subclasses ($R_{ij} \geq 0.05$)

Some findings can be identified from Figure 2 as follows. (1) Those subclasses of Section H have close intra-section relatedness with one another and broad cross-section relatedness with the subclasses from other sections. The entire network contains some star-type subnets and most of their core nodes are the subclasses of Section H, such as H01L, H01M, H02K, and H02G. "*Section H: Electricity*" is proved to be a TF with a high degree of interdisciplinarity and the subclasses of

Section H are also interdisciplinary subfields that have a broad range of relatedness with other fields and subfields.

(2) Various subclasses from the class of H04 (red nodes), coupled with several subclasses of Section G (blue nodes), construct a cluster that is notable in the entire network, as shown in the bottom right of Figure 2. The red nodes in the cluster have two distinctive characteristics. The first is that these nodes have higher frequency in our sample, and the second is that these red nodes in the cluster have close intra-class relatedness. They also have some cross-section relatedness with the subclasses of Section G, but the relatedness is weaker than intra-class relatedness. Generally, the structure of the cluster represents the technological structure of class H04. It is also confirmed that Class H04 has strong relatedness with "*Section G: Physics*".

(3) Subclasses of various classes have different characteristics of relatedness. As described, the subclasses of Class H04 are mainly co-classified with the subclasses of Class H04 and Section G, which form a relatively isolated cluster. Subclasses of Class H01 and Class H02 show different characteristics compared with H04. Those subclasses have widespread distribution in the entire network. Their frequencies are far less than those of the subclasses of Class H04, but they have much broader relatedness, especially cross-section relatedness.

As shown in Figure 2, most nodes from other sections in the network are connected by the nodes of H01 and H02. By contrast, the subclasses of Class H03 construct a loose cluster (see bottom of Figure 2), which mainly concentrates on intra-class relatedness, with the subclass of H03K as the core of the cluster. Class H05 has few subclasses and most of them are located on the edge of the network, which means that they are less active in the field of electricity.

(4) Besides the aforementioned subclasses, other active subclasses, such as H01L, H01M, H02K, H02G, H02H, and H02K, have a higher degree of interdisciplinarity. They have relatively broad relatedness with many other fields and subfields.

4.3 Monitoring Trends

Relatedness analysis based on co-classification not only can display the structure of a field but can also be used to monitor the trend of technological diversification [17]. This study chose a number of high-frequency subclasses (IPC codes) to construct a co-classification network and to display it visually so that the trend in the TF of electricity can be determined. Only codes whose frequencies were ranked in the top 30 in the sample were chosen to represent the hot domains of Section H. Their co-classified relations were displayed on the co-classification network map. Those hot domains characterized by a high degree of interdisciplinarity may involve some emerging TFs, driven by the knowledge relatedness and characterized by technological diversification [18]. Therefore, such a co-classification network could indicate the trend to a certain extent. The threshold of $R_{ij} \geqq 0.05$ was also introduced into the visualization to clearly preserve the high-strength relatedness.

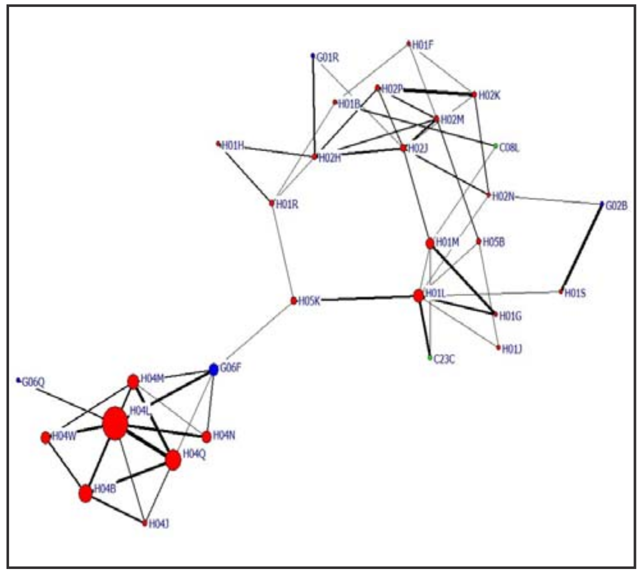

Figure 3. Visualization of high-frequency subclasses and their relatedness ($R_{ij} \geq 0.05$)

Some findings can be identified from Figure 3 as follows.

(1) The entire network is divided into two subnets obviously. One subnet is composed by some subclasses of Class H04 (on the lower left side of Figure 3) while the other is composed of a large number of subclasses of Class H01 and Class H02 (on the upper right side of Figure 3). The two subnets demonstrate different structures. The first subnet contains fewer nodes but their frequencies are much larger and they have a stronger relatedness than the nodes of the second subnet. As shown in Figure 3, the second subnet covers a larger number of nodes but their frequencies and relatedness strengths cannot catch up with those of the first subnet.

(2) Among the top 30 frequent subclasses in Figure 3, 6 do not belong to Section H, including G06F, G02B, C08L, G01R, C23C, and G06Q. All of them are from Sections G and C. Evidently, electricity has a great deal of knowledge and technology in common with sections of physics, chemistry, and metallurgy. Those subfields in the three sections have a significant influence on the evolving technological trends in electricity.

(3) The two subnets are connected by a single line that is considered as a bridge from the perspective of SNA. The two nodes of H05K and G06F are gatekeepers. "*Bridge*" and "*gatekeeper*" are terms used to describe the roles of some critical actors and the relations between them [19]. In the social network, the bridge and the gatekeeper are considered to be much more powerful in

influencing and controlling the process of information dissemination and knowledge transfer. According to such a concept, the two subclasses of H05K and G06F play critical roles in enhancing the knowledge linkages and spillovers among different subfields, as well as the technological relatedness and diversification in the field of electricity.

(4) Several notable pairs or clusters have stronger relatedness, such as H04L–H04B–H04Q–H04M–H04W, H01S–G02B, H02P–H02K, and H01M–H01. Those subfields characterized by a high degree of interdisciplinary and close relatedness may represent the direction of technological diversification in the TF of electricity. Furthermore, the pairs and clusters with highly-degree interdisciplinary and close relatedness have a significant likelihood of giving birth to new technologies.

5. Conclusion

This study provided a picture of the interdisciplinary status of the TF of electricity through the statistics and mapping of IPC codes from SIPO. The analytical results enabled us to identify the focus and trends, as well as to visualize and measure the relatedness among various technological domains. Through the frequency statistic on 4-digit IPC codes, we found that "H04L: Transmission of digital information" is the most active subclass that can be seen as the focus, accounting for nearly a quarter of patent applications. In addition, subclasses, such as H04W, H04N, H04Q, and H01L were also active and constituted a secondary core area. Through the co-classification analysis, we confirmed that electricity is indeed a field characterized by a high degree of interdisciplinarity and has a broad range of technological relatedness with many fields and subfields.

Relatedness is the driver of technological diversification and convergence; thus, the interdisciplinary domains composed of subclasses with characteristics of high frequency as well as intensified and broad relatedness with other subclasses, such as H04L–H04B–H04Q–H04M–H04W, H01S–G02B, H02P–H02K, and H01M–H01, can be regarded as the breeding ground for technological diversification and convergence. These interdisciplinary domains are significantly inclined to give birth to new technologies.

Acknowledgment

The study was supported by the National Social Science Foundation of China (No.14CTQ018).

References

[1] Joo, S. H., Kim, Y. (2010). Measuring relatedness between technological fields. *Scientometrics*, 83 (2), 435-454.

[2] Yoon, J., Park, Y., Kim, M., Lee, J., Lee, D. (2014). Tracing evolving trends in printed electronics using patent information. *Journal of Nanoparticle Research*,16 (7), 2471.

[3] Kim, M. S., Kim, C. (2012). On a patent analysis method for technological convergence. *Procedia Social and Behavioral Sciences,* (40) 657-663.

[4] Choi, J. Y., Jeong, S., Kim, K. (2015). A study on diffusion pattern of technology convergence: patent analysis for Korea. *Sustainability*, 7 (9), 11546-11569.

[5] Kim, E., Kim, J., Koh, J. (2014). Convergence in Information and Communication Technology (ICT) Using Paten Analysis. *Journal of Information Systems and Technology Management*, 11 (1), 53-64.

[6] Ko, N., Yoon, J., Seo, W. (2013). Analyzing interdisciplinary of technology fusion using knowledge flows of patents. *Expert Systems with Applications*, 41 (4) 1955-1963.

[7] Ko, S. S., Ko, N., Kim, D., Park, H., Yoon, J. (2014). Analyzing technology impact networks for R&D planning using patents: combined application of network approaches. *Scientometrics*, 101 (1), 917-936.

[8] Curran, C. S., Leker, J. (2011).Patent indicators for monitoring convergence — examples from NFF and ICT. T*echnological Forecasting & Social Change*, 78 (2), 256-273.

[9] Su, H. N. (2012). Visualization of global science and technology policy research structure. J*ournal of the American Society for Information Science and Technology,* 63 (2) 242-255.

[10] Choi, S., Kim, H., Yoon, J., Kim, K., Lee, J. Y. (2013). An SAO-based text-mining approach for technology roadmapping using patent information. R&D Management, 43(1), 52-74.

[11] Park, S., Lee, S. J., Jun, S. (2015). A network analysis model for selecting sustainable technology. *Sustainability,* 7 (10), 13126-13141.

[12] Noruzi, A., Abdekhoda, M. (2012). Mapping Iranian patents based on International Patent Classification(IPC), from 1976 to 2011. *Scientometrics*, 93 (3), 847-856.

[13] Park, H., Yoon, J. (2014). Assessing coreness and intermediarity of technology sectors using patent co-classification analysis: the case of Korean national R&D. *Scientometrics,* 98 (2), 853-890.

[14] SIPO. (2015). Instruction and guide of SIPO database. Retrieved April 20, 2015, at: http://www.sipo.gov.cn/zljs/ .

[15] Archambault, E., Larivière, V. (2011). Scientific publications and patenting by companies: A study of the whole population of Canadian firms over 25 years. *Science and Public Policy,* 38 (4) 269-278.

[16] Leydesdorff, L. (2008). Patent classifications as indicators of intellectual organization. Journal of the American Society for Information Science and

Technology, 59 (10), 1582-1597.

[17] Koren, M., Tenreyro, S. (2013). Technological diversification. *American Economic Review*, 103 (1), 378-414.

[18] Hacklin, F., Wallin, M. W. (2013).Convergence and interdisciplinarity in innovation management: a review, critique, and future directions. *Service Industries Journal,* 33 (7-8), 774-778.

[19] Otte, E., Rousseau, R. (2001). Social Network Analysis: a powerful strategy, also for the Information Sciences. *Journal of Information Science*, 28 (6), 441-453.

APPENDIX

For those IPC codes the paper refers to, some codes' titles are not showed in the text and maps, they are specially listed in the appendix.

• G06-Computing; Calculating; Counting

• G05-Controlling; Regulating

• G01-Measuing; Testing

• H05K-Printed circuits; Castings or constructional details of electric apparatus; Manufacture of assemblages of electrical components

• H05B-Electric heating; Electric lighting not otherwise provided for

• H04J-Multiplex communication

• H02P-Control or regulation of electric motors, generators, or dynamo

• H02N-Electric machines not otherwise provided for

• H02M-Apparatus for conversion between AC and AC, between AC and DC, or between DC and DC, and for use with mains or similar power supply systems; Conversion of DC or AS input power into surge output power; Control or regulation thereof

• H02K-Dynamo

• H02H-Emergency protective circuit arrangements

• H02G-Installation of electric cables or lines, or of combined optical and electric cable or lines

• H01S-Devices using stimulated emission

• H01R-Electrically

• H01J-Electic discharge tubes or discharge lamps

• H01H-Electric switches; Relays; Selectors; Emergency protective devices

• H01G-Capacitors; Capacitors, rectifiers, detectors, detectors, switching devices, light

• ·H01F-Magenets; Inductances; Transformers; Selection of material for their magnetic properties

• H01B-Cables; Conductors; Insulators; Selection of materials for their conductive, insulating, or dielectric properties

• G06Q-Data processing systems or methods, specially adapted for administrative, commercial, financial, managerial, supervisory or forecasting purposes; Systems or methods specially adapted for administrative, commercial, financial, managerial, supervisory or forecasting purposes, not otherwise provided for

• G06F-Electric digital data processing

• G02B-Optical elements, systems, or apparatus

• G01R-Measuing electric variables; Measuring magnetic variables

• C23C-Coating metallic material; Coating material with metallic material by diffusion into the surface, by chemical conversion or substitution; Coating by vacuum evaporation, by sputtering, by ion implantation or by chemical vapor deposition, in general

• C08L-Compositions of macromolecular compounds