An Empirical Study on Utilizing Pre-grant Publications in Patent Classification Analysis

Chung-Huei Kuan[1] and Chan-Yi Lin[2]

[1] *maxkuan@mail.ntust.edu.tw*
National Taiwan University of Science and Technology, Graduate Institute of Patent Research, No. 43 Sec. 4 Keelung Rd., Taipei City, 106 (Taiwan, R.O.C.)

[2] *u9703220@gmail.com*
National Taiwan University of Science and Technology, Graduate Institute of Patent Research, No. 43 Sec. 4 Keelung Rd., Taipei City, 106 (Taiwan, R.O.C.)

Abstract

Patent classification analyses are usually conducted using issued patents. Issued patents however suffer lengthy examination and the derived analytic results reflect R&D activities occurring considerable time in the past. The only option for an analyst to reduce such observational time delay is to use the so-called pre-grant publications (PGPubs) that are open to public 18 months after patent applications are filed. The PGPubs and their corresponding issued patents are both assigned classification symbols. If the two sets of symbols are very different, using patent classification analysis on PGPubs to observe R&D activities is dubious. This study therefore compares the United States Patent Classification (USPC) symbols assigned to about 235,000 pairs of U.S. utility patents issued in 2012 and their PGPubs in three ways, each corresponding to an approach of a conventional patent classification analysis: (1) considering only the class codes of the main classification symbols; (2) considering only the main classification symbols; and (3) considering both main and auxiliary classification symbols. The study finds that only the class codes of the PGPub main classification symbols are reliable enough for patent classification analysis as there are about 78% of the PGPubs have identical class codes as their corresponding issued patents.

Conference Topic

Patent analysis

Introduction

A patent application is classified during its prosecution process based on its inventive content by an examiner and one or more classification symbols are assigned in accordance with a standard scheme such as International Patent Classification (IPC), Cooperative Patent Classification (CPC), U.S. Patent Classification (USPC), etc. Patent classification analysis (PCA) is a popular practice by patent analysts using the patent classification symbols, and it is so popular that, to the authors' knowledge, all commercial patent analytic systems/services, such as Thomson Innovation® and WIPS Global®, have various types of PCA built-in.

A common type of PCA is to investigate the R&D focuses of an entity (i.e., a company, an institute, a country, a technical field, etc.). An analyst gathers the patents affiliated with the entity, collects the classification symbols assigned to these patents, counts the number of times each classification symbol is assigned to these patents, and usually produces a diagram such as a histogram, a heat map, etc., to visually manifest the assignment frequencies of the classification symbols. By observing the diagram, the analyst then claims that the entity has its R&D focused in a few technical areas denoted by the most frequently assigned classification symbols.

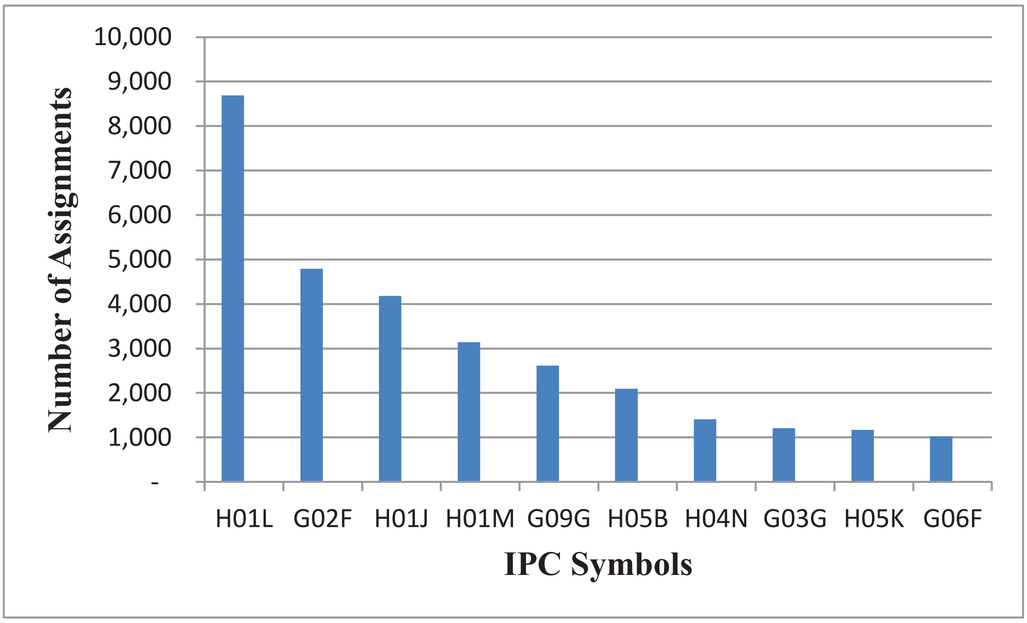

Figure 1. A sample histogram from a fictitious PCA.

A sample histogram from a fictitious PCA using IPC symbols for a company is shown in Figure 1. As illustrated, the company is considered to have its R&D effort mainly focused in the field Semiconductor Devices denoted by the most frequently assigned IPC symbol H01L.

Other than the real-life application described above, patent classification symbols are considered as a viable source of technological information by researchers, and various types of PCA have been proposed in the literature. To mention just a few, the number of different classification symbols assigned to an entity's patents is used as a proxy to the entity's technological diversity (cf. Lerner, 1994), the co-classification of patents (i.e., patents assigned one or more identical classification symbols) is used to investigate the linkage among technologies (cf. OECD, 1994), or the relationships among organizations (cf. Leydesdorff, 2008). There are also studies investigating the technological relatedness of two entities using the classification symbols assigned to their patents (cf. Jaffe, 1986; 1989). In addition, the classification symbols of a patent's forward and backward citations are used to evaluate the patent's "generality" and "originality" (cf. Henderson, Jaffe, & Trajtenberg, 1997). However it should be noted that there are opinions considering the existing patent classification schemes are "never intended to provide conceptual delineations of technology areas, but instead identify inventions by function at very low levels of abstraction in order to serve as aids to prior art searching" (Allison et al., 2004).

As described above, PCA can be used to observe the focus of an entity's R&D activities up to the time of analysis or, if the entity's latest patents are gathered, of the entity's recent R&D activities. However, what is revealed by the latter is actually not the R&D activities happened around the time of analysis but a considerable amount of time in the past. To see this, the curve with diamond marks in Figure 2 depicts the distribution of U.S. utility patents issued in the year 2012 according to their application years. About three quarters of the 2012-issued utility patents are actually filed between 2007 and 2010. In other words, if a histogram similar to Figure 1 is derived from these 2012-issued patents, the revealed R&D focuses actually occur and disperse in a period of time quite in the past.

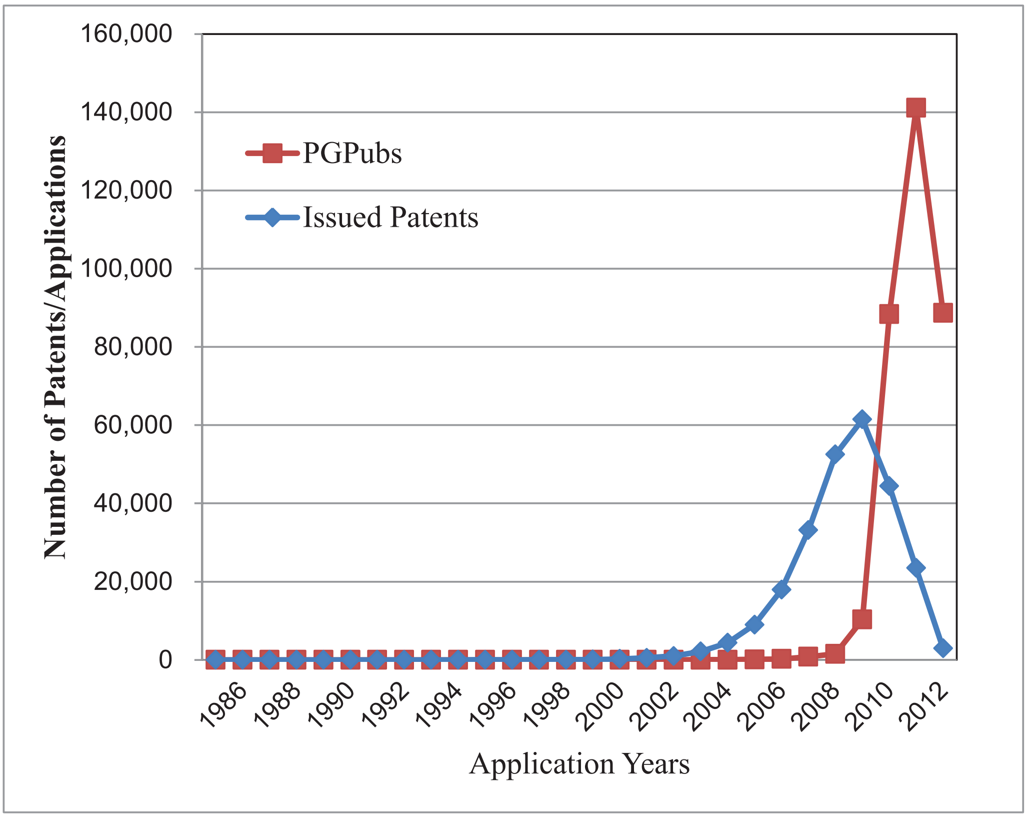

Figure 2. Distributions of 2012 issued patents and PGPubs based on application years.

The only possible way to reduce this time delay is to use the so-called *pre-grant publications* (PGPubs), instead of the issued patents. A patent application usually undergoes an early publication process before the patent is issued by the authority or before the patent application is given up by the applicant. Again taking Figure 2 as example, the curve with square marks depicts the distribution of U.S. PGPubs published in the year 2012 according to their application years. As illustrated most PGPubs are filed between 2010 and 2012, which are concentrated in a more limited period of time and in a more recent past.

The early publication process is a common practice for authorities across various nations and regions. For example, U.S. Patent Act (35 U.S.C. § 122(b)) specifies that, "each application for a patent shall be published … promptly after the expiration of a period of 18 months from the earliest filing date for which a benefit is sought under this title." There are indeed exceptions that an application is not early published if the application is (i) no longer pending; (ii) subject to a secrecy order; (iii) a provisional application; (iv) an application for a design patent; or (v) requested by the applicant. These exceptions are not common and, for utility patent applications, which are the most common type of patent applications, it is very possible that an issued utility patent is early published. According to our statistics, there are 253,580 utility patents issued in the year 2012 and 17,993 of them (7.1%) do not have corresponding PGPubs.

When a patent application is filed, the patent application is initially classified and classification symbols are assigned so as to route the patent application to an appropriate examiner team (USPTO, 2004). Then, after the patent application has undergone substantive examination, its examiner may alter the initial classification and assign different classification symbols (USPTO, 2005). As such a PGPub and its subsequently issued patent have their respective classification symbols and their classification symbols may not be identical.

PCAs usually utilize the issued patents, instead of PGPubs, most likely due to that the PGPubs have not undergone substantive examination, and their classification symbols may not fully reflect their inventive contents. Yet PGPubs are better subjects for investigating the latest R&D focuses as they do not suffer lengthy pendency and strict screening by the examination process as reflected in Figure 2.

This study therefore tries to investigate the adequacy of using PGPub classification symbols for PCA. If the answer is yes, analysts can effectively reduce the time delay of their analytic observations to about 18 months, which is a significant improvement. On the other hand, even if the answer is no, analysts would know that PGPub classification symbols are not reliable, and they should avoid using them or at least be cautious about PCAs based on PGPub classification symbols.

Methodology

To investigate the adequacy of PGPub classification symbols for PCA, we collected U.S. utility patents issued in 2012 and their corresponding PGPubs for comparison. Utility patent is chosen because, for the three types of U.S. patents, utility patent is the most common and numerous one, design patents do not undergo the early publication process, and there are only a small number of plant patents. According to our statistics, there are only 868 plant patent applications filed each year between 1992 and 2011 on the average.

Each U.S. utility patent/PGPub is classified with three classification schemes: IPC, CPC, and USPC, and we choose the USPC symbols for comparison. This is because USPC is the default scheme for United States Patent and Trademark Office (USPTO) (USPTO, 2012), the IPC symbols are most likely machine-converted from the USPC symbols, and the CPC are not popular yet. Most importantly, USPC scheme does not have versions as it is updated every two months and the USPC symbols of all documents contained in USPTO databases are thoroughly and automatically re-classified accordingly (Wolter, 2012). In other words, when the USPC symbols of an issued patent are compared against those of its PGPub, whether the USPC symbols are of the same version is not an issue. One may question that USPC, as a domestic scheme, may not be representative. However, we believe that what this study observes from using U.S. patents and USPC could provide us at least some hint when dealing with patents of different countries and using different classification schemes.

Like all other classification schemes, USPC provides a hierarchical taxonomy of technical areas. Each USPC symbol contains a class code and a subclass code separated by "/." For example, a USPC symbol 623/2.1 has class code 623 and subclass code 2.1. The class code (e.g., 623) represents a highest level of non-overlapping technical area whereas the subclass code (e.g., 2.1) represents a lower level of technical area belonging to the one denoted by the class code. For subclass codes under the same class code, they may have hierarchical relationship among themselves. For example, 623/2.11 and 623/2.12 represent parallel technical areas but the two technical areas both belong to the technical area denoted by the symbol 623/2.1 (USPTO, 2012).

A U.S. utility patent/PGPub is assigned one or more USPC symbols. Among them, one and only one is expressed in boldface in the patent/PGPub documents. For issued patents, the official name for the bold-faced symbols is *original classification* symbols and, for the normal-faced symbols, *cross-reference classification* symbols by USPTO. As to PGPubs, the official name for the bold-faced ones is *primary classification* symbols and, for the normal-faced ones, *secondary classification* symbols. For simplicity's sake, we refer to the bold-faced symbols as the *main classification* symbols whereas the rest of the normal-faced symbols as the *auxiliary classification* symbols, whether or not they are from issued patents or PGPubs. The main classification covers the novel and non-obvious information contained in a patent/PGPub whereas the auxiliary classification covers other information considered to be valuable for searching (USPTO, 2012).

To determine whether PGPub classification symbols is adequate for PCA, we use the classification symbols assigned to the corresponding issued patents as reference as they are assigned by examiners after substantive examinations and therefore assumed to have better reflected the inventive contents of the patents.

Table 1 provides a number of examples where the sets of classification symbols assigned to three U.S. utility patents issued on 2015/02/10 and their PGPubs are listed side by side for comparison. As illustrated in Table 1, the two sets of classification symbols may not be identical, and the set assigned to the issued patent indeed seems to be more detailed than that assigned to the corresponding PGPub.

Table 1. The classification symbols assigned to three sample pairs of PGPubs/patents.

PGPub no./Patent no.	*PGPub symbols*	*Patent symbols*
20140289912/8,955,161	**850/18**	**850/1**; 250/339.11; 250/339.14; 73/105; 850/5; 850/50; 850/6
20120124680/8,955,160	**726/34**	**726/34**
20110252484/8,955,159	**726/32**	**726/32**; 380/201; 705/57; 726/27; 726/31; 726/33

There are quite some researches involving the measurement of similarity between nodes in a hierarchical taxonomy of concepts, which can be applied to classification symbols as well. For example, in one so-called edge-based approach, the similarity between two nodes is calculated based on the numbers of edges from the root of the hierarchical structure to the two nodes and to their nearest common ancestor node (Slimani, Yagahlane, & Mellouli, 2008). Similar edge-based approaches can be found in McNamee (2013). There are also so called node-based approaches, which capture a node's feature in the hierarchical structure as a vector and calculate a similarity measure based on the concept vectors of two nodes (cf. Liu, Bao, & Xu, 2012).

These studies do have their academic merit but cannot directly tell us whether PGPub classification symbols is reliable or not for PCA. We therefore adopt a different and practical treatment to the comparison of the classification symbols. First, we notice that existing commercial analytic systems/services conduct PCA using one of three simple approaches:

- PCA using Approach 1 counts only the class codes of the patent or PGPub main classification symbols so as to obtain a broad picture of the distribution of R&D activities;
- PCA using Approach 2 counts only the main classification symbols and ignores all auxiliary classification symbols of patents or PGPubs, considering that the main classification symbols are the most representative ones; and
- PCA using Approach 3 counts all patent or PGPub classification symbols with no distinction between main and auxiliary classification symbols, believing all classification symbols are equally important.

To demonstrate the three approaches, using the Patent Symbols column listed in Table 1 as example:

- Approach 1 counts the class codes 850 as being assigned once, 726 being assigned twice;
- Approach 2 counts each of the main classification symbols 850/1, 726/34, and 726/32 as being assigned once; and
- Approach 3 counts each of the 14 classification symbols as being assigned once.

Please note that, to the authors' knowledge, commercial analytic systems/services ignore the hierarchical relationship between classification symbols. For the above example, 850/6 is actually a technology area belonging to that of 850/5 but commercial analytic systems conducting PCA using Approach 3 treat 850/5 and 850/6 as denoting distinct technology areas probably for simplicity's sake.

Then, to see whether PCA using one of the above approaches on PGPubs classification symbols would deliver trustworthy result, we conduct three analyses as follows, each corresponding to one of the approaches above:

- Analysis 1 compares the main classification class codes of PGPubs to those of the corresponding issued patents.
- Analysis 2 compares the main classification symbols of PGPubs to those of the corresponding issued patents and calculates the consistency rate.
- Analysis 3 compares the sets of classification symbols of PGPubs to those of the corresponding issued patents.

Then all three analyses calculate the percentage of PGPubs having *identical* main classification class codes, main classification symbols, and sets of classification symbols to their corresponding issued patents. Since commercial analytic systems/services ignore the hierarchical relationship between classification symbols, our three analyses follow the same practice.

A 100% percentage indicates that PCA on PGPubs using one of the approaches would yield a result identical to that using their issued patents, meaning that using PGPubs can achieve reduced time delay with total accuracy. But a 0% percentage implies that PCA on PGPubs using one of the approaches delivers totally incorrect result. We therefore specifically refer to the percentage as *consistency rate* so as to avoid confusion with the general term *percentage*.

If statistically there is a very high consistency rate or similarity from the PGPubs, a histogram such as Figure 1 obtained from PGPubs using Approach 1, 2, or 3 would be very close to one from the corresponding subsequently issued patents. An analyst then can confidently utilize the PGPubs for PCA by Approach 1, 2, or 3 and achieve a reduced time delay.

To demonstrate the three analyses, again using the three sample pairs of PGPubs/patents listed in Table 1 as example:

- Analysis 1 shows that PCA using Approach 1 on PGPubs has a 100% consistency rate (i.e., all three pairs' PGPubs have identical main classification class codes to those of their issued patents);
- Analysis 2 shows that PCA using Approach 2 on PGPubs has a 66% consistency rate (i.e., except the first pair, the other two pairs' PGPubs have identical main classification symbols to those of their issued patents); and
- Analysis 3 shows that PCA suing Approach 3 on PGPubs has a 33% consistency rate (i.e., only the second pair's PGPub has an identical set of classification symbols to that of its issued patent).

For PCA using Approach 3, the simple consistency rate described above is too narrow to give us a complete picture. For example, even though the two sets of classification symbols from the third pair of patent/PGPub listed in Table 1 are different, the PGPub classification symbol {726/32} is actually a proper subset of the issued patent's classification symbols {726/32, 380/201, 705/57, 726/27, 726/31, 726/33} and therefore still captures a portion of the inventive content. The calculation of the consistency rate however ignores this condition.

Therefore in conducting Analysis 3, we divide the PGPub-patent pairs into 5 categories based on the relationships between their sets of classification symbols so as to gain more insight.

- Category 1: their sets of classification symbols are identical (i.e., $\{\text{PGPub}\} = \{\text{Patent}\}$).
- Category 2: their sets of classification symbols are entirely different (i.e., $\{\text{PGPub}\} \neq \{\text{Patent}\}$ and $\{\text{PGPub}\} \cap \{\text{Patent}\} = \emptyset$).
- Category 3: the PGPub's set of classification symbols is a proper subset of that of the corresponding patent (i.e.,$\{\text{PGPub}\} \neq \{\text{Patent}\}$ and $\{\text{PGPub}\} \subset \{\text{Patent}\}$).
- Category 4: the patent's set of classification symbols is a proper subset of that of the corresponding PGPub (i.e.,$\{\text{PGPub}\} \neq \{\text{Patent}\}$ and $\{\text{Patent}\} \subset \{\text{PGPub}\}$).

- Category 5: their sets of classification symbols are not entirely different, do not belong to each other, and have a non-empty intersection (i.e.,$\{\mathrm{PGPub}\} \neq \{\mathrm{Patent}\}$,$\{\mathrm{PGPub}\} \not\subset \{\mathrm{Patent}\}$, $\{\mathrm{Patent}\} \not\subset \{\mathrm{PGPub}\}$, and $\{\mathrm{Patent}\} \cap \{\mathrm{PGPub}\} \neq \emptyset$).

Then, for the patent/PGPub pairs belonging to each category, we calculate an average Jaccard Coefficient (Jaccard, 1901) as expressed in (1) where {PGPub} and {Patent} are the two sets of classification symbols assigned to the PGPub and the corresponding issued patent, respectively. Jaccard Coefficient, or Jaccard Index, or Jaccard Similarity Coefficient, was originally designed for comparing similarity between sample sets, and has already been applied in patent bibliometrics such as co-citation analysis (Small, 1973). Here we use it to capture the degree of discrepancy between {PGPub} and {Patent}.

$$\mathrm{J} = \frac{|\{\mathrm{PGPub}\} \cap \{\mathrm{Patent}\}|}{|\{\mathrm{PGPub}\} \cup \{\mathrm{Patent}\}|} \tag{1}$$

Findings

We collected 253,580 utility patents issued in the year 2012 from USPTO database. After removing those having no corresponding PGPub, those having no classification symbol (e.g., these patents are withdrawn and withdrawn patents do not have patent classification symbols recorded in the USPTO database), and for unknown reason those having no main classification symbols, there are total 234,966 patents eligible for analysis. As mentioned in the previous section, USPC is updated every two months and all patents are re-classified accordingly. We collected the USPC symbols assigned to the 234,966 patents and their corresponding PGPubs under the USPC scheme up to 2013/10/31.

An initial statistics shows that the 234,966 patents have average 3.9 USPC symbols and their corresponding PGPubs have average 2.2 USPC symbols, and that 64.16% of the 234,966 patents have a greater number of USPC symbols than that of the corresponding PGPubs, indicating that issued patents seem do have more careful assignment of classification symbols than their PGPub counterparts. In some extreme cases, PGPub No. 2010/0316607 has the greatest number of USPC symbols (48) among all PGPubs whereas patent No. 8,179,540 has the greatest number of USPC symbols (65) among all patents. The latter is also the case having the greatest difference (63) between the issued patent and the corresponding PGPub.

Analysis 1

For each pair of the 234,966 PGPubs and corresponding issued patents, we compared the class code of the PGPub's main classification symbols against that of the corresponding issued patent, and we found that the consistency rate is 77.89%. That is, 183,024 out of the 234,966 pairs of PGPubs and patents have identical main classification class codes, and the remaining 51,942 pairs (22.11%) have difference main classification class codes. In other words, there is a 22.11% probability that a PGPub's main classification class code does not accurately reflect the inventive content of the corresponding patent.

Analysis 2

For each pair of the 234,966 PGPubs and corresponding patents, we compared the main classification symbol of the PGPub against that of the corresponding issued patent, and we found that the consistency rate drops to only 36.42%. That is, 85,584 out of the 234,966 pairs of PGPubs and patents have identical main classification symbols, and the rest 149,382 pairs (63.58%) have different main classification symbols. In other words, there is a very significant 63.58% probability that a PGPub's main classification symbol does not accurately reflect the inventive content of the corresponding patent.

Analysis 3

For the 234,966 pairs of PGPubs and corresponding patents, we categorized them into 5 categories based on the relationships between their sets of classification symbols, and calculated the average Jaccard Coefficient for each category. The result is summarized in Table 2.

Table 2. Comparison result from Analysis 3.

Category	*Pairs*	*Percentage*	*Avg. Jaccard Coefficient*	*Std. Deviation*
1	14,958	6.37%	1	0
2	89,981	38.30%	0	0
3	63,057	26.84%	0.34	0.16
4	10,693	4.55%	0.45	0.15
5	56,277	23.95%	0.22	0.11

As illustrated, PGPubs in Category 1 are those having identical sets of classification symbols to their issued patents and their share (6.37%) among the 234,966 PGPubs is exactly the consistency rate of Analysis 3.

PGPubs in Category 2 are those having totally different sets of classification symbols from their issued patents and, for a PCA on these Category-2 PGPubs using Approach 3, the analytic result would be totally incorrect, but PGPubs of this category has the greatest share (about 38%) among all PGPubs.

PGPubs in Category 3 are those having sets of classification symbols being proper subsets to those of their issued patents, and cover about 27% of all PGPubs. For these Category-3 PGPubs, their classification symbols capture only 34% of the inventive content as reflected by their average Jaccard Coefficient. We can imagine that, for a PCA on Category-3 PGPubs using Approach 3, a histogram such as Fig. 1 would miss a significant amount of information.

Category 4 is a special case where PGPubs have sets of classification symbols that are proper supersets to those of the corresponding issued patents, and therefore covers the smallest share (less than 5%). For these Category-4 PGPubs, their classification symbols capture all inventive content but unfortunately provide on the average 55% (1-0.45) surplus and erroneous information. Again we can imagine that a histogram from PCA on Category-4 PGPubs using Approach 3 would contain too much noise.

Category 5 is a combination of Categories 3 and 4, meaning these 24% of the PGPubs have sets of classification symbols that not only miss significant amount of information but also provide significant amount of erroneous information, as reflected by the very limited average Jaccard Coefficient (0.22).

Conclusion

This study arises out of an attempt to use PGPub classification symbols for PCA so as to investigate an entity's latest R&D focuses with limited time delay. It is however speculated that the PGPub classification symbols are not carefully assigned and their adequacy for PCA has to be determined first.

We therefore gathered 234,966 pairs of issued patents and corresponding PGPubs, and compared their classification symbols in accordance with the three approaches that a commercial patent analytic system/service usually employ.

Assuming that the classification symbols of the corresponding issued patents better reflect the inventive contents of the patents and as such using them as reference, we find that, if the commercial patent analytic systems/services count the main classification symbols, or the entire sets of classification symbols of the PGPubs for PCA, only 36.42% of the PGPubs have identical main classification symbols, and only 6.37% of the PGPubs have identical sets of classification symbols to those of the corresponding issued patents. PCA using PGPubs as described can hardly be considered as reliable.

The best candidate for using PGPubs in PCA is the PGPubs' main classification class codes. We find that as high as 77.89% of the PGPubs have identical main classification class codes to those of the corresponding issued patents. The main classification class codes, however, represent the broadest technical areas and using them to investigate R&D focuses would provide only limited insight.

This study can be further carried out as follows. In order to make the main classification class codes even more useful for PCA, the consistency rate for each individual class can be determined. For some classes that have statistically very high consistency rate, PGPubs assigned with these class codes can be used for PCA with high confidence whereas, for classes of low consistency rate, an analyst should avoiding using them for PCA.

Additionally, one may be curious about why some class codes reveal higher consistency rates than the others. We speculate that, for some well-developed technical fields, the consistency rates of their class codes would be high as the classification of the related technology should be familiar to the examiners whereas for emerging technical fields, the consistency rates of their class codes would be low as the examiners may have different opinions on what the related technology should be classified. The investigation of this speculation is currently under way.

If both reduced time delay and better analytic insight are required, an analyst would require a better tool that can take the hierarchical relationship among classification symbols into consideration. If this kind of tool is available, we speculate that some specific technical areas may reveal a high consistency rate or similarity measure even for PCA using Approaches 2 and 3. The identification of these specific technical areas and how reliable the PGPub classification symbols are in these specific technical areas can be further investigated.

Acknowledgments

This study is funded by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant No. MOST 103-2221-E-011-115.

References

Allison, J.R., Lemley, M.A., Moore, K.A., & Trunkey, R.D., (2004), Valuable Patents. Georgetown Law Journal, 92, 435–479.

Henderson, R.M., Jaffe, A., & Trajtenberg, M., (1997), University versus Corporate Patents: A Window on the Basicness of Invention. *Economics of Innovation and New Technology, 5*(1), 19–50.

Jaccard, P., (1901), Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 547–579.

Jaffe, A.B., (1986), Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits and market value. *The American Economic Review, 76*(5), 984–1001.

Jaffe, A.B., (1989), Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy, 18*(2), 87–97.

Leydesdorff, L., (2008), Patent Classifications as Indicators of Intellectual Organization. *Journal of the American Society for Information Science and Technology, 59*(10), 1582–1597.

Lerner, J., (1994), The Importance of Patent Scope: An Empirical Analysis. *RAND Journal of Economics, 25*(2), 319–333.

Liu, H.Z., Bao, H., & Xu, D., (2012), Concept Vector for Similarity Measurement Based on Hierarchical Domain Structure. *Computing and Informatics, 30*(5), 881–900.

McNamee, R.C., (2013), Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example. *Research Policy, 42*(4), 855–873.

OECD (1994), The Measurement of Scientific and Technological Activities Using Patent Data as Science and Technology Indicators: Patent Manual. Paris: OECD Publishing. DOI: 10.1787/9789264065574-en.

Slimani, T., Yagahlane, B.B., and Mellouli, K., (2008), A new similarity measure based on edge counting. *Proceedings of the World Academy of Science, engineering and Technology, 23*, 773–777.

Small, H., (1973), Co‐citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science, 24*(4), 265–269.

USPTO (2004), Pre-Grant Publicaiton (PGPub) Global Concept of Operations. USPTO. Available on-line at http://www.uspto.gov/web/offices/dcom/olia/aipa/PGPubConOps.pdf.

USPTO (2005), Handbook of Classification. USPTO. Available on-line at http://www.uspto.gov/web/offices/opc/documents/handbook.pdf.

USPTO (2012), Overview of the U.S. Patent Classification System (USPC). USPTO. Available on-line at http://www.uspto.gov/patents/resources/classification/overview.pdf.

Wolter, B., (2012), It takes all kinds to make a world–Some thoughts on the use of classification in patent searching. *World Patent Information, 34*(1), 8–18.