A Study of the Correlation Between Number of Classification Symbols and Patent Citation Count

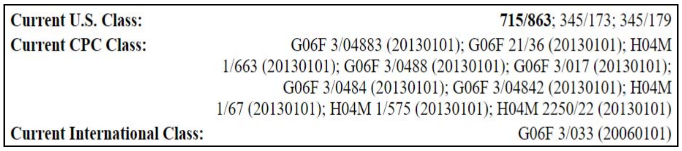
Chung-Huei Kuan, Tsai-Hsuan Yang
Graduate Institute of Patent, National Taiwan University of Science and Technology, Taipei, Taiwan R.O.C.

***Abstract*—Patents' classification symbols are a valuable source of information for patent analysis. The number of different classification symbols assigned to a patent is often considered as an indicator to the patent's technical scope, breadth, or diversity. However, the validity of the indicator is dubious. This study speculates that, if this indicator indeed reflects some characteristics of a patent, and that, if more different classification symbols a patent is assigned with, the patent is considered more valuable or desirable, the indicator should have a positive correlation with the patent's citation count, which is widely accepted as representative of the patent's quality, value, importance, or impact. Using empirical data and statistical analysis, this study finds that, for patents of three different ages, their numbers of Cooperative Patent Classification symbols at two different levels are all positively correlated to their citation counts, confirming the validity of this simple indicator. This finding is especially helpful when evaluating young patents that are issued for only a limited period of time.**

I. Introduction

Every patent includes one or more classification symbols as part of its bibliometric data. These symbols are assigned during the patent's application process by authority according to the patent's disclosed invention and a standard scheme such as International Patent Classification (IPC), Cooperative Patent Classification (CPC), U.S. Patent Classification (USPC), etc. For example, Fig. 1 is a partial screen capture of U.S. Patent No. 7,657,849 from USPTO (United States Patent and Trademark Office) full-text database. As illustrated, the patent is assigned with symbols from USPC ("Current U.S. Class"), CPC ("Current CPC Class"), and IPC ("Current International Class").

Patents' classification symbols are a valuable source of information as they are determined by professional personnel of the authority, and are representative of the patents' technical contents. A common type of patent classification analysis is to investigate the R&D focuses of an entity (i.e., a firm, an institution, a country, etc.) by observing the assignment frequencies of the classification symbols of its patents, usually in the form of a histogram.



Current U.S. Class:	**715/863**; 345/173; 345/179
Current CPC Class:	G06F 3/04883 (20130101); G06F 21/36 (20130101); H04M 1/663 (20130101); G06F 3/0488 (20130101); G06F 3/017 (20130101); G06F 3/0484 (20130101); G06F 3/04842 (20130101); H04M 1/67 (20130101); H04M 1/575 (20130101); H04M 2250/22 (20130101)
Current International Class:	G06F 3/033 (20060101)

Fig. 1. A partial screen capture of U.S. Patent No. 7,657,849 from USPTO

Fig. 2 is one such histogram from a patent classification analysis using IPC symbols rounded to the 3rd level (i.e., the first 4 digits). Based on the diagram, the entity is considered as having its R&D effort mainly focused in the field "Semiconductor Devices" denoted by the most frequently assigned IPC symbol "H01L."

Various approaches of utilizing patent classification symbols have also been proposed in the literature. For example, Henderson, Jaffe, and Trajtenberg [8] used Herfindahl-Hirschman Index (HHI) to see how concentrated or dispersed the classification symbols of a patent's forward and backward citations are, and interpreted the result as the patent's "generality" and "originality." Schmoch, et al [25] considered that if two classification symbols have high co-assignment frequency (for example, the 4-digit CPC symbols G06F and H01M shown in Fig. 1 are co-assigned to a patent), the technical fields denoted by the classification symbols should be more related. In other words, the co-assignment frequencies among classification symbols are used to investigate the linkage among technologies. Jaffe [11][12] and Leydesdorff [18] used the classification symbols assigned to organizations' patent portfolios to investigate the relationships among organizations. McNamee [20] measured the similarity between patents based on classification symbols' structural information in the hierarchical classification scheme using the so-called Jaffe Distance. Breschi, Lissoni, and Malerba [2] used classification symbols to see how firms diversify their innovative activities across knowledge-related technological fields. Leydesdorff, Kushnir, and Rafols [19] used patent classification symbols to represent the technology space as a network map of vertices (technology fields) and weighted edges (distance between the technology fields).

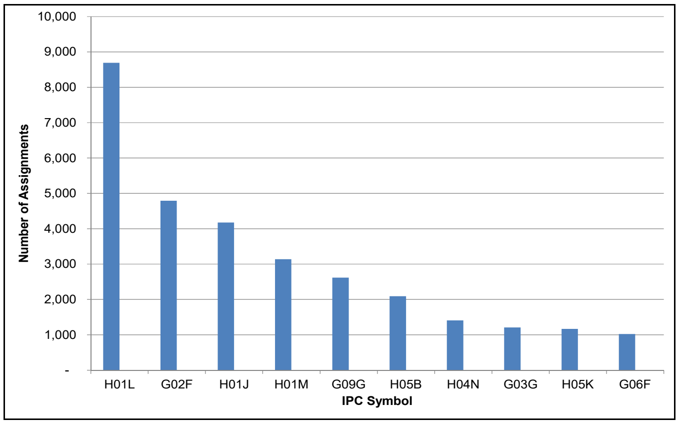


Fig. 2. A histogram from a fictitious patent classification analysis

This study is funded by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant No. MOST 105-2221-E-011-105.

978-1-890843-36-6 ©2017 PICMET

II. Patent Classification Scheme

All classification schemes provide a hierarchical taxonomy of technical areas, and each node (i.e., technical area) of the structure is associated with a unique classification symbol. For example, a IPC symbol "E02B 3/04" represents a 5th-level technical area subordinate to a 4th-level one denoted by the symbol "E02B 3/00" or "E02B 3," which in turn is subordinate to a 3rd-level one denoted by "E02B," as illustrated in Fig. 3, which is a partial IPC classification scheme under the section symbol "E."

As technologies advance, the 5-level structure of IPC is no longer enough and technical areas of even deeper levels are defined. For example, the IPC symbol "E02B 3/14" looks like a 5th-level one but is actually two levels deeper beneath the symbol "E02B 3/04," as illustrated in Fig. 4, which is a partial IPC classification scheme under the 4th-level symbol "E02B 3/00."

This scenario applies to all classification schemes. Therefore, two classification symbols may look different in appearance, but they may represent technical areas having hierarchical or superordinate/subordinate relationship, instead of two distinct technical areas. In other words, the degree of difference between the areas denoted by "E02B 3/04" and "E02B 3/14" should be smaller than that between the areas denoted by "E02B 3/04" and "E02B 3/16."

In addition, the classification symbols assigned to patents may have different relevance to their inventive contents. Using Fig. 1 as example, the boldface USPC symbol "715/863" covers the novel and non-obvious part of the patent 7,657,849, whereas the rest of the symbols expressed in normal face covers other part of the patent's inventive content considered to be valuable for searching [26]. All classification schemes allow the specification of at least one boldfaced or main symbols, and none or more normal-faced auxiliary symbols.

III. Patent Technology Scope

Patent Technology Scope (PTS) is a simple indicator that counts the number of distinct main and auxiliary classification symbols assigned to a patent as a proxy to the patent's breadth or diversity. If a patent is assigned 5 different classification symbols, its PTS is 5.

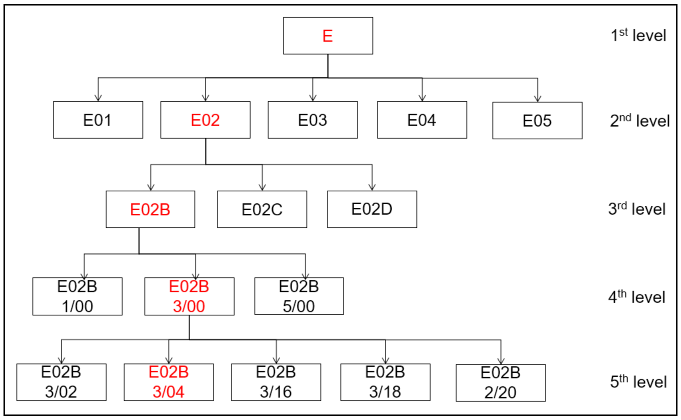


Fig. 3. A partial IPC classification scheme under the section "E"

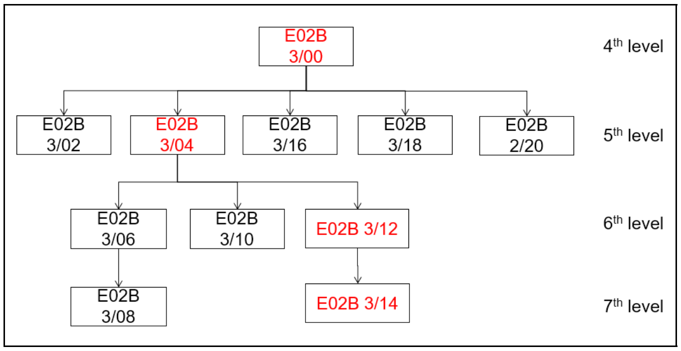


Fig. 4. A partial IPC classification scheme under the symbol "E02B 3/00"

Lerner [14][15][16] is probably the first researcher who used the number of distinct classification symbols as a measure of a patent's scope ("breadth of patent protection"). By applying this simple indicator to biotechnology firms, Lerner claimed that the performance or value of these firms is significantly affected by their PTSes. The name Technology Scope was coined by Ernst [4], who considered that PTS reflects the diversity and therefore the technological quality of patent applications, and the indicator may be used to monitor competitors' patenting activities. Su, Chen, and Lee [23] observed litigated and non-litigated patents and found that litigated patents, usually considered as more valuable ones, have higher average numbers of IPC and USPC symbols than those of the non-litigated ones.

A number of related researches seem to follow a theme that the classification symbols are first mapped to more abstract "technology fields," and it is the number of "technology fields" that is counted. Garcia-Vega [5] grouped 2nd-level IPC symbols (i.e., the first 3 digits) into 49 fields, and measured 544 European firms' technological diversification among these fields, and claimed that there is a statistically significant positive relationship between technological diversity and innovation at the firm level. Similarly, Ozman [22] investigated the breadth ("the range of different subjects that a technology field draws upon") and depth ("the extent to which a certain field is exploited in detail") of 30 technology fields and 40 largest firms in biotechnology and telecommunications using patents' IPC symbols. The author claimed that these technology fields and firms are largely scattered in terms of breadth and depth, and that the field biotechnology has the highest breadth and depth. Leten, Belderbos, and Looy [17] assigned each IPC symbol to one of 30 different technology fields. Then, the technology class information of the patents in a firm's patent portfolio was used to derive technological diversification, which is defined as the spread of the patent portfolio over technology classes.

There are also researches that, instead using simple counts of symbols or "technology fields," an index is used to measure how concentrated or disperse patent portfolios are. Chen, Jang, and Wen [3] used HHI on IPC symbols of 73 Taiwanese IC design firms and concluded that this measure indeed reflects the spread or distribution amongst technology classes of a company's current technology portfolio. More recently, Hu and Rousseau [10] combined PTS with *h*-index [9] to propose a number of new patent indicators. For example, the proposed

IPCh index is determined similarly to the *h*-index but, instead of using patent citation counts, the authors used patents' PTSes based on 3rd-level IPC symbols. Then, if an entity is said to have IPCh index equal to 4, the entity has at least 4 patents, each assigned with at least 4 IPC symbols, similar to the interpretation of *h*-index.

As can be seen from the above brief literature review, PTS, usually counting symbols rounded to a higher level and either used directly or indirectly through mapping to technology fields, is considered equivalent to the technological diversity of an entity. Despite the seemingly wide acceptance of PTS, its validity is not without challenge. Allison et al. [1] considered that counting classification symbols is dubious as 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."

We also have reservations towards this simple indicator as it possesses some dubious behavior. First of all, PTS is highly dependent on the classification scheme used. Fig. 5 is a partial screen capture of U.S. Patent No. 8,776,261 from USPTO full-text database. As illustrated, this patent has PTS equal to be 1, regardless of the USPC, CPC, or IPC classification scheme used. In contrast, for the U.S. patent 7,657,849 shown in Fig. 1, its PTS varies significantly and may be equal to 3, 10, or 1 if the classification scheme used is USPC, CPC, or IPC.

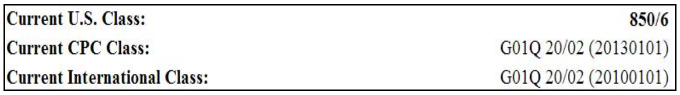
Secondly, PTS is also dependent on which level the classification symbols are rounded to. Again taking Fig. 1 as example, if the CPC symbols are considered to the more coarse sub-class level (i.e., the 3rd level), the patent's PTS is 2 as there are two distinct symbols "G06F" and "H04M" whereas, if the CPC symbols are considered to the finer sub-class level (i.e., the 4th level and the digits before the "/"), the patent's PTS becomes 4 as there are 4 distinct symbols "G06F 3," "G06F 21," "H04M 1," and "H04M 2250".

Finally and most importantly, by counting each symbol as 1, PTS implicitly assumes that each symbol represents a technical area of identical "size" or "breadth." This practice is especially questionable when the symbols are not at the same level and/or when some of the symbols have superordinate/subordinate relationship. For example, if the above-mentioned symbols "E02B 3/04" and "E02B 3/14" are co-assigned to a same patent, counting them as 2 ignores the reality that "E02B 3/14" is subordinate to "E02B 3/04" and therefore denotes a technical area which is a subset of the one denoted by "E02B 3/04".

Intrigued by these doubts, we intend to conduct a more thorough investigation into the validity of PTS by employing a large amount of empirical data.

IV. METHODOLOGY

A patent's citation count is the number of times the patent is referenced as relevant prior art by applicants or examiners of subsequent patent applications and has long been accepted as an indication to the patent's value, impact, or importance (cf. [7][13][24]).



Current U.S. Class:	**850/6**
Current CPC Class:	G01Q 20/02 (20130101)
Current International Class:	G01Q 20/02 (20100101)

Fig. 5. A partial screen capture of U.S. Patent No. 8,776,261 from USPTO

The patent citation count is actually one of the earliest patent bibliometric indicators after Narin [21] pointed out its significant similarity to paper citation count in his pioneering work.

Therefore, instead of verifying whether classification symbols at a same level represent equally broad technical areas, or how classification symbols at different levels should be normalized, we speculate that, if PTS indeed reflects some nice characteristics of a patent, be it the scope, breadth, or diversity of the patent, and that, if more distinct classification symbols a patent is assigned with, the patent is considered more valuable or desirable, PTS should have a positive correlation with the patent's citation count.

We as such decided to use U.S. utility patents issued in the years 2007, 2009, and 2011 and collected their main and auxiliary CPC symbols and citation counts up to Dec. 31st, 2013 so as to evaluate their correlation. Patent data from three different years are used because patent citation counts need time to accumulate. According to Hall, Jaffe, and Trajtenberg [6], patents are most frequently cited after they are issued for 5 years, and the citation counts drop after 7 years. By counting citations up to the end of 2013, these patents have accumulated citations for average 2, 4, and 6 years, respectively. We, therefore, are able to observe how PTSes correlate patent citation counts for patents beginning to pick up citations, patents that are moderately cited, and patents within their citation peaks. We may also gain more insight by observing how the correlation differs for patents of different ages.

A U.S. utility patent is assigned with USPC, CPC, and IPC symbols as shown in Figs. 1 and 5. We choose CPC symbols because (1) USPTO has given up USPC and switched to use CPC as the default classification scheme after June 2015; (2) even though currently only USPTO and EPO (European Patent Office) are using CPC, we expect that it will replace IPC and become the standard scheme for authorities around the world; and (3) even though most prior works used IPC symbols, CPC is an extension to IPC and has pretty much identical 1st- to 4th-level structure as IPC does.

A final decision is about which level the CPC symbols should be rounded to. Most prior works chose to round the classification symbols to the 3rd level without giving a reason. However, by rounding symbols to a same level, the prior works actually avoided the issue of two classification symbols having hierarchical relationship and that they should not be counted like they represent two distinct technical areas. We also speculate that the 3rd level is chosen because it is not too coarse and not too fine either. Currently, IPC and CPC have about 130 2nd-level symbols, about 630 3rd-level symbols, and about 7,400 4th-level symbols[1].

[1] A statistics may be found at: http://www.wipo.int/classifications/ipc/en/ITsupport/Version20160101/transformations/stats.html.

In this study, we observe the correlation between patent citation counts and PTSes counting CPC symbols rounded not only to the 3rd level but also to the 4th level, so as to see whether the two approaches would differ in terms their correlation with the patent citation counts.

A summary of the empirical data is listed in Table I. As shown, patents from year 2007 have the greatest average citation count as it is accumulated over a longer period of time. The year 2007 also has the greatest standard deviation to the average citation count as the greater influence of some patents is better manifested over time. The average PTSes, their standard deviations, and the maximum PTSes, whether from counting symbols to the 3rd level or to the 4th level, are all pretty much the same across all three years despite the significant differences among the numbers of patents, which is quite interesting but not unreasonable.

V. RESULT

To gain an overview of how citation counts are related to PTSes, we first calculated the average citation counts from patents of various PTSes and plot the results in Figs. 6 and 7 with the PTSes along the horizontal axis and the corresponding average citation counts along the vertical axis.

As shown in Fig. 6, for patents whose PTSes are counted with symbols rounded to the 3rd level, the average citation counts indeed increase for patents with PTSes from 1 to 7, even though not very obvious. When PTSes are greater than 7, there are dramatic fluctuations in the average citation counts. The curves of Fig. 7 reveal similar behavior. For patents whose PTSes are counted from symbols rounded to the 4th level, the average citation counts increase incrementally for patents with PTSes from 1 to 9. There are similar fluctuations for patents whose PTSes are greater than 9.

The fluctuations are resulted from a handful of patents having especially large or small citation counts among a very small portion (less than 1%) of patents whose PTSes are greater than 7 (Fig. 6) or 9 (Fig. 7). For example, in Fig. 6, there are only two patents with PTS 16 and another two with PTS 20 in year 2007. For the former, one is cited 8 times and the other is not cited at all, leading to average citation count 4. As to the latter, the two patents are cited 45 and 3 times, respectively, leading to average citation count 24.

We then combined the average citation counts corresponding to the PTSes from 1 to 9 of the three years into Table II. For example, the "3rd-level" columns list the average citation counts for patents from the three years, respectively, corresponding to PTSes counted from 3rd level CPC symbols. Please note, for these columns, the average citation counts for PTSes 8 and 9 (i.e., greater than 7) are omitted because these numbers are distorted as shown in Fig. 6 and explained above.

TABLE I. SUMMARY OF EMPIRICAL DATA

	2007	2009	2011
No. of patents	137,720	152,280	208,124
Avg. citations	5.06	2.67	0.98
Avg. citation std. dev.	10.73	6.39	2.62
Avg. PTS, 3rd level	1.84	1.78	1.82
Avg. PTS std. dev., 3rd level	1.17	1.13	1.18
Max. PTS, 3rd level	22	21	22
Avg. PTS, 4th level	3.00	2.93	3.04
Avg. PTS std. dev., 4th level	2.45	2.40	2.46
Max. PTS, 4th level	70	77	80

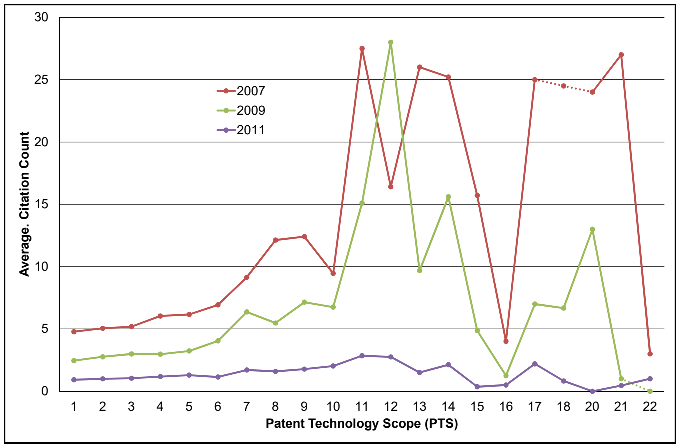


Fig. 6. Average citation counts vs. PTSes from 3rd-level CPC symbols

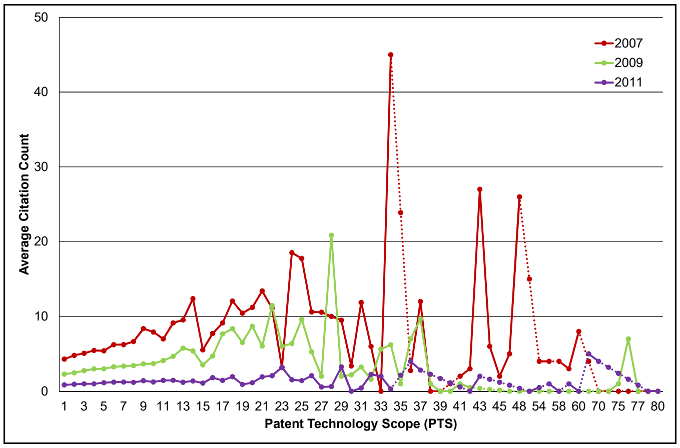


Fig. 7. Average citation counts vs. PTSes from 4th-level CPC symbols

From Table II we can see that, for all three years and for PTSes using 3rd-level and 4th-level symbols, the average citation counts indeed have a positive relationship with the PTSes. However, we can also see that, for patents with close PTSes, their difference is only obscurely manifested by the average citation counts. For example, for patents with PTSes 1 and 2, or 2 and 3, their average citation counts have very small differences, even for patents that have aged.

TABLE II. AVERAGE CITATION COUNTS FOR PTSES FROM 1 TO 9

PTS	2007		2009		2011	
	3rd level	4th level	3rd level	4th level	3rd level	4th level
1	4.78	4.29	2.45	2.28	0.92	0.85
2	5.05	4.79	2.76	2.46	0.99	0.93
3	5.18	5.08	2.99	2.75	1.05	0.99
4	6.04	5.45	2.97	2.99	1.18	1.00
5	6.16	5.41	3.23	3.00	1.29	1.14
6	6.93	6.23	4.04	3.27	1.14	1.21
7	9.14	6.23	6.36	3.36	1.71	1.23
8	-	6.64	-	3.43	-	1.19
9	-	8.37	-	3.65	-	1.40

But for patents with more distant PTSes, such as those with PTSes 3 and 6, their differences are more obvious in terms of the average citation counts. We can also see that PTSes using 3rd-level symbols are more discriminant as evident from their average citation counts spanning greater ranges than those using 4th-level symbols. This is reasonable as 3rd-level symbols represent larger and more distinct technical areas than those represented by the 4th-level symbols. Patents with greater PTSes using 3rd-level symbols, therefore, should have broader contents.

In addition to the above visual observations, we have calculated Pearson's correlation coefficients between the PTSes and the citation counts for all patents from the three years. The result is summarized in Table III. As shown, patents' PTSes are indeed positively and significantly correlated with their citation counts in all three years. The significant fluctuations from patents of large PTSes (e.g., greater than 7 or 9) observed in Figs. 6 and 7 have little influence as these patents only account for no more 1% of all patents.

VI. CONCLUSION

From the analysis reported in the last section, we have confirmed that the simple indicator PTS indeed captures some characteristics of patents. However, due to the limited correlation coefficient values (e.g., the highest one in Table III is only 0.077), the practical application of PTS should be cautious.

We think that, given two patents with different PTSes, whether the one with greater PTS is really a more valuable patent than the one with lower PTS should be considered along with the following factors.

First of all, for patents with PTSes greater than 7 or 9, depending on the level which their symbols are rounded to, there is actually not enough statistical evidence to support that higher PTSes imply more citation counts due to too few samples.

In addition, PTSes obtained from using 3rd-level symbols should be more trustworthy than those obtained from using 4th-level symbols.

TABLE III. PEARSON'S CORRELATION COEFFICIENTS

	2007	2009	2011
For PTS using 3rd-level symbols	0.053**	0.059**	0.037**
For PTS suing 4th-level symbols	0.077**	0.071**	0.047**

**p-value less than 0.01 significance level

Furthermore, due to the limited correlation coefficient values, the two patents' PTSes should have a greater difference (e.g, one is 6 and one is 2) so that a more confident conclusion may be drawn.

Our observation would be especially helpful when evaluating young patents that are issued for only a limited period of time. As they are too young to accumulate meaningful citation counts, and as such they cannot be differentiated reliably using citation counts, the indicator PTS may be employed to fill the gap.

REFERENCES

[1] J. R. Allison, M. A. Lemley, K. A. Moore, and R. D. Trunkey, "Valuable Patents," Georgetown Law J., vol. 92, no. 3, pp. 435–1309, 2004.

[2] S. Breschi, F. Lissoni, and F. Malerba, "Knowledge-relatedness in firm technological diversification," Res. Policy, vol. 32, pp. 69–87, 2003.

[3] J. H. Chen, S.-L. Jang, and S. H. Wen, "Measuring technological diversification: identifying the effects of patent scale and patent scope," Scientometrics, vol. 84, no. 1, pp. 265–275, 2010.

[4] H. Ernst, "Patent information for strategic technology management," World Patent Inf., vol. 25, no. 3, pp. 233–242, 2003.

[5] M. Garcia-Vega, "Does technological diversification promote innovation? An empirical analysis for European firms," Res. Policy, vol. 35, no. 2, pp. 230–246, 2006.

[6] B. H. Hall, A. B. Jaffe, and M. Trajtenberg, "The NBER patent citation data file: lessons, insights and methodological tools," No. w8498, National Bureau of Economic Research, 2001.

[7] B. H. Hall, A. B. Jaffe, and M. Trajtenberg, "Market value and patent citations," Rand J. Econ., vol. 36, no. 1, pp. 16–38, 2005.

[8] R. M. Henderson, A. B. Jaffe and M. Trajtenberg, "University versus corporate patents: a window on the basicness of invention," Econ. Innovation New Tech., vol. 5, pp. 19–50, 1997.

[9] J. E. Hirsch, "An index to quantify an individual's scientific research output," Proceedings of the National Academy of Sciences of United States of America, vol. 102, pp. 16569–16572, 2005.

[10] X. Hu, and R. Rousseau, "A simple approach to describe a company's innovative activities and their technological breadth," Scientometrics, vol. 102, no. 2, pp. 1401–1411, 2015.

[11] A. B. Jaffe, "Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value," Am. Econ. Rev., vol. 76, pp. 984–1001, 1986.

[12] A. B. Jaffe, "Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers," Res. Policy, vol. 18, pp. 87–97, 1989

[13] A. B. Jaffe, M. S. Fogarty, and B. A. Banks, "Evidence from patents and patent citations on the impact of NASA and other federal labs on commercial innovation," J. Ind. Econ., vol. 46, no. 2, pp. 183–206, 1998.

[14] J. Lerner, "The impact of patent scope: an empirical examination of new biotechnology firms," CSIA Discussion Paper 91-4, Kennedy School of Government, Harvard University, 1991.

[15] J. Lerner, "The importance of patent scope: an empirical analysis," Rand J. Econ., pp. 319–333, 1994.

[16] J. Lerner, and R. P. Merges, "Patent scope and emerging industries: biotechnology, software and beyond," Competing in the age of digital convergence, Harvard Business Press, 1997.

[17] B. Leten, R. Belderbos, and B. V. Looy, "Technological diversification, coherence, and performance of firms," J. Prod. Innovat. Manag., vol. 24, no. 6, pp. 567–579, 2007.

[18] L. Leydesdorff, "Patent classifications as indicators of intellectual organization," J. Am. Soc. Inf. Sci. Tec., vol. 59, pp. 1582–1597, 2008.

[19] L., Leydesdorff, D. Kushnir, and I. Rafols, "Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC)," Scientometrics, vol. 98, pp. 1583–1599, 2015.

[20] R. C. McNamee, "Can't see the forest for the leaves: similarity and distance measures for hierarchical taxonomies with a patent classification example." Res. Policy, vol. 42, pp. 855–873, 2013.

[21] F. Narin, "Patent bibliometrics," Scientometrics, vol. 30, no. 1, pp. 147–155, 1994.

[22] M. Ozman, "Breadth and depth of main technology fields: an empirical investigation using patent data," Science and Technology Policies Research Centre, Working Paper Series 7.01, 2007.

[23] H. N. Su, C. M.-L. Chen, and P.-C. Lee, "Patent litigation precaution method: analyzing characteristics of US litigated and non-litigated patents from 1976 to 2010," Scientometrics, vol. 92, no. 1, pp. 181–195, 2012.

[24] M. Trajtenberg, "A penny for your quotes: patent citations and the value of innovations," Rand J. Econ., vol. 21, pp. 172–187, 1990.

[25] U. Schmoch, F. Laville, M. Pianta, and G. Sirilli, "The measurement of scientific and technological activities: using patent data as science and technology indicators," Patent Manual, 1994.

[26] United States Patent and Trademark Office, Overview of the U.S. Patent Classification System (USPC), 2012. Retrieved from http://www.uspto.gov/patents/resources/classification/overview.pdf.