Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 147 (2014) 153 – 159

Procedia
Social and Behavioral Sciences

ICININFO

Comparative Analysis of Patent Citations of Different Fields: In Consideration of the Data Size Dependency of Statistical Measures

Fuyuki Yoshikane[a,b,]*

[a]University of Tsukuba, Faculty of Library, Information and Media Science, 1-2, Kasuga, Tsukuba 305-8550, Japan
[b]University of Tsukuba, Research Center for Knowledge Communities, 1-2, Kasuga, Tsukuba 305-8550, Japan

Abstract

This study focused on the classifications assigned to patents, and examined the number of different classifications of patents citing each patent, namely the diversity of citing fields. Specifically, we observed the expected values for growth in the average number of citing fields according to the increase in the data size (i.e., the cumulative number of citations) when regarding the observation period as being in a synchronic state and assuming that the strength of potential connections between each subject patent and citing fields is constant. A total of 347,327 patent applications published in Japan in 1993 were subjected to the analysis. The results of comparisons between fields showed that, irrespective of the cumulative number of citations, the transition of the diversity of citing fields is stable with higher values in the fields of "performing operations; transporting," "chemistry; metallurgy," and "textiles; paper" (sections B, C, and D of the International Patent Classification), while it is stable with lower values in the field of "fixed constructions" (section E). Moreover, as for the field of electricity (section H), it was found that the growth rate of the diversity of citing fields is markedly slow—in other words, a patent in this field tends to receive citations repeatedly from the same fields as before and to be hardly ever cited by patents belonging to fields different from them even when the cumulative number of citations increases.

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.

Keywords: Bibliometrics; Scientometrics; Citation analysis; Patent; Japan.

* Corresponding author. Tel.: +08-29-859-1346.
E-mail address: fuyuki@slis.tsukuba.ac.jp

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.
doi:10.1016/j.sbspro.2014.07.132

1. Introduction

Looking at citations in the analysis of patents is meaningful in the sense that it allows us to understand the position and importance of a patent as part of "the continuity, lineage, or stream of technology," rather than its "patentability" or "economic value." This is because citations between patents are considered to express, not the benefits that the cited patents themselves generate in the market, but the technological impact on subsequent patents directly or indirectly linked with them. Lee (2009) posits different types of patent value: one is direct economic value, which can be apprehended through patent licensing or income from royalties; and another is technological value, which can be measured by looking at the number of patent citations. In addition, many studies treat citations as indications of the relationship, diffusion, origin, and general background in technology and knowledge (e.g., Meyer, 2000; Jang, Lo, & Chang, 2009; Haruna, Jinji, & Zhang, 2010; Li-Ying et al., 2013).

While there is a large body of research analyzing the overall structure of citation relations for patents or for the countries to which patents belong—namely the socio-centered network—(e.g., Li et al., 2007), there has been very little research seeking to clarify the characteristics of fields relating to the diffusion of technology through observing the citation network surrounding each patent—namely the ego-centered network (Wasserman & Faust, 1994)—individually and tracing its growth. It is considered that the significance of the approach taken in this study, wherein ego-centered networks are dealt with, lies in that it allows us to map trends in the changes relating to connections of individual patents, which cannot be captured solely by monitoring the citation network as a whole. This study clarifies the characteristics of technological fields regarding the quantitative growth of ego-centered citation networks of individual patents, focusing on the citing classifications, that is, based on the diversity of fields of patents that cite the subject patents.

2. Data

Information sources in this study were the NTCIR test collections compiled by the National Institute of Informatics (NII), Japan, and we used the full text of the "patent gazette (publication of unexamined patent applications)" published in Japan; the 3,496,253 documents published in the ten years between 1993 and 2002 from NTCIR-7 Patent Mining Test Collection (Nanba et al., 2008); and the 1,757,361 documents published in the five years between 2003 and 2007 from NTCIR-8 Patent Translation Test Collection (Fujii et al., 2010). Approximately 350,000 documents were published in each of these years. The publication number, the application number, the publication date, the application date, the classifications assigned according to the International Patent Classification (IPC), and the publication and application numbers of cited patents were extracted from each patent record.

Table 1. Basic quantities regarding the subject patents.

	P	TL	CL	N	$CL_{citing}(N)$
A: human necessities	27980	630.88	1.84	24520	2.60
B: performing operations; transporting	82825	592.91	2.05	74271	2.71
C: chemistry; metallurgy	51616	600.73	2.26	74133	3.45
D: textiles; paper	7081	604.60	2.11	7132	3.19
E: fixed constructions	13825	606.19	1.79	7295	2.18
F: mechanical engineering, etc.	34007	589.66	1.78	21274	2.33
G: physics	120529	596.47	1.72	122666	2.42
H: electricity	107636	588.85	1.74	75739	2.34

A total of 347,327 patent applications published in 1993 were subjected to the analysis. We investigated patents that cite the subject patents during fifteen years following their publication, that is, from 1993 to 2007, and identify the classifications assigned to those patents. Table 1 sets out the basic quantities related to the patents published in 1993 belonging to each field. The aggregation of patents for each field has been based on the section level, which

is the top layer in IPC (WIPO, 2010). *P* is the number of the subject patents; *TL* is the average value for the time lag between the dates of application and publication; *CL* is the average value for the number of different classifications assigned at the subclass level (the fourth layer in IPC); and *N* is the overall number of citations the subject patents received from 1993 to 2007. Because it is common to assign multiple classifications to a patent, the sum of patents (*P*) under the sections from A to H exceeds the total number of the subject patents, i.e., 347,327.

$CL_{citing}(N)$ is the number of different classifications at the subclass level assigned to patents that cite the subject patent and thus represents the diversity of fields in forward citations, that is to say, how many fields citations are made by. In other words, $CL_{citing}(N)$ corresponds to the average value of the outdegree per patent (excluding isolated nodes) in the citation network where the ties are oriented from each subject patent to classifications citing it. Figure 1 shows examples of the citation networks for patents. These are directed bipartite graphs that take as their nodes each subject patent and the classifications of the patents citing it. The codes contained within the diagrams—for instance, A23K—represent classifications at the subclass level of IPC. In the left diagram, patent (a), for example, is cited by patents belonging to B01D (separation) or B02C (crushing). The right diagram shows the conditions in which some citations have occurred in addition (i.e., the cumulative number of citations has increased). Regarding patent (a), for example, the degree in the citation network (i.e., the number of different classifications citing it) grows from 2 to 3 according to the increase in the cumulative number of citations. Here, not only might the number of citations by the same classifications increase, but also citations by another classification—that is, A23K (feeding-stuffs)—have been added. After all, the number of citing classifications for a patent depends on the cumulative number of citations and thus $CL_{citing}(N)$ is expressed as a function of the cumulative number of citations.

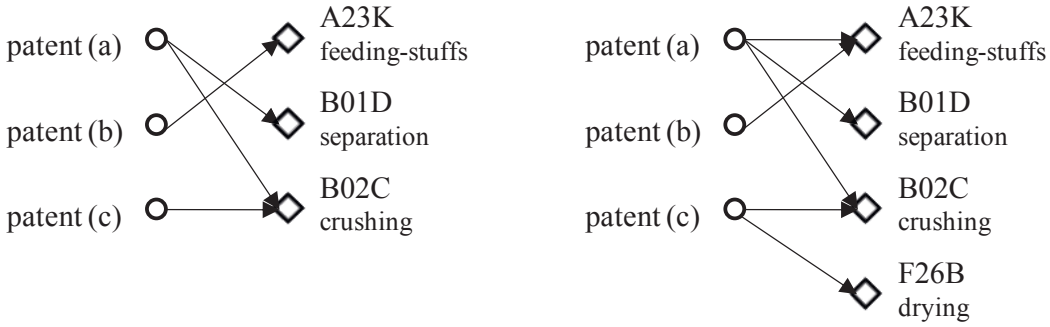

Fig. 1. Bipartite graphs representing the citation network for patents.

3. Methodology

Through tracing citations each of the subject patents receives, we observed the change in the value of $CL_{citing}(n)$, that is, the average value per patent of the number of different citing classifications. On the basis of the result of observation, the characteristics for eight fields (sections of IPC) were examined. As stated in the previous section, $CL_{citing}(n)$ corresponds to the average value of the degree in the ego-centered citation networks of individual patents.

Specifically, we observed the expected values for growth in the average number of citing classifications, $CL_{citing}(n)$, according to the increase in the data size (i.e., the cumulative number of citations, *n*) when regarding the period from 1993 to 2007 as being in a synchronic state and assuming that the strength of potential connections between each subject patent and citing classifications is constant. Here, considering "the strength of connections between the subject patents and citing classifications" to be constant means assuming a probability model in which "the population probability that a citation of each patent by each classification occurs" does not change.

The methods used for observing the change in the value of an index according to the change in the size of data—in other words, the data size dependency of an index—are as follows: (1) interpolation and extrapolation that estimate expected values for the number of occurring events on the basis of a binomial distribution and its

Poisson approximation (e.g., Good & Toulmin, 1956; Efron & Thisted, 1976), and (2) Monte-Carlo simulation that calculates average values for an arbitrary index through repeated random sub-sampling (e.g., Tweedie & Baayen, 1998; Baayen, 2001). Although the former allows us to observe changes in the number of occurring events beyond the range of the original data size, it can be applied only to limited indices and it is difficult to be applied to complex phenomena such as the growth of networks. On the other hand, the latter is applicable to arbitrary indices and has been used also for exploring the growth of the degree of nodes in networks (Yoshikane & Kageura, 2004).

In this study, for observing the growth of the citation networks, we carried out Monte-Carlo simulation in which we performed 1,000 random sub-samplings for each one-twentieth interval of the original data size N (i.e., $n = N/20, 2N/20, 3N/20, ..., N$) for each of sections A–H. For each data size n, we calculated the average value of 1,000 trials for $CL_{citing}(n)$ (i.e., $CL_{citing}(N/20)$, $CL_{citing}(2N/20)$, $CL_{citing}(3N/20)$, ..., $CL_{citing}(N)$). Moreover, we plotted the 95% Monte-Carlo confidence interval for $CL_{citing}(n)$ at each data size n. The value and growth rate of $CL_{citing}(n)$ were compared among sections.

4. Results and discussion

As of 1993, the early publication system, in which, upon request, application documents are published even before the lapse of one and a half years, had yet to be introduced in Japan. As shown by Table 1, the average value for the time lag from application to publication, TL, is somewhat longer than one and a half years. When a patent is cited, there are not only cases where its publication number is referenced but also those where its application number is referenced. This study takes into account both types of citations. Figure 2 shows the cited count by year for the patents published in 1993. It illustrates the transitions of the numbers of citations: citations referencing the publication number, citations referencing the application number, and all citations in which the two types are summed up. From the figure, we can first see that, in most citations just after publication, it is not the publication number but the application number that is being referenced. Because of the publication time lag, the fact that a patent is being cited in patents published in the same year or the following year basically means that the patent must be cited by the application number before the publication number is assigned to it. It can be supposed that many of these are self-citations by applicants (applicant firms) of the cited patents.

The publication time lag is also considered to influence the initial peak of the number of all citations, which is located in three years after publication (i.e., in 1996). The number of all citations shows weak obsolescence after this peak, and then, having stopped falling in around 2002, it greatly increases in 2004. In Japan, to list prior art documents in patent application became obligatory in September 2002 (Sato & Iwayama, 2006). This would contribute to the increase in the number of citations after 2002. Following this, the number of all citations peaks for the second time in 2005, after which it begins to decrease again. The total number of citations that the subject patents have received during the fifteen years subsequent to their publication is 299,271.

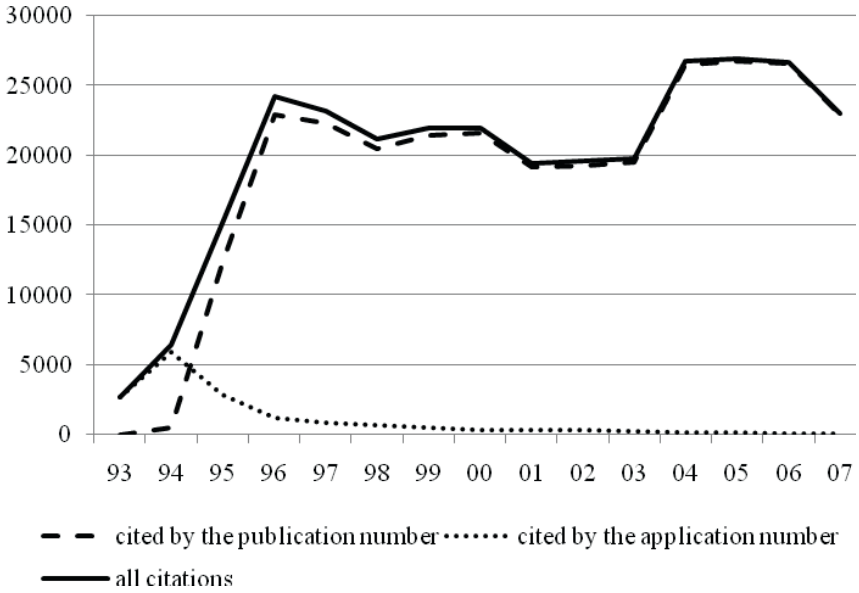

Fig. 2. The number of patents citing the subject patents for each year.

Expected values for growth in the average number of citing classifications, $CL_{citing}(n)$, according to the increase in the scale of data were calculated through Monte-Carlo simulation on the basis of repeated random sub-sampling for each section. Figure 3 shows the developmental profiles of $CL_{citing}(n)$. Through this figure, it can be grasped how citing fields become diverse as the cumulative number of citations increases when the strength of potential connections between each subject patent and citing classifications is constant.

As for sections B (performing operations; transporting), C (chemistry; metallurgy), and D (textiles; paper), $CL_{citing}(n)$ shows transitions that are stable with higher values irrespective of the scale of data. In particular, the values of $CL_{citing}(n)$ are remarkably higher in sections C and D, which are chemistry (materials chemistry)-related fields, than in the other sections. In contrast, $CL_{citing}(n)$ in section E (fixed constructions) shows a transition that is stable with the lowest values. Thus, we can say that, while each patent belonging to the former (i.e., sections B, C, and D) tends to receive citations from a diverse range of fields, that belonging to the latter (i.e., section E) tends to receive citations from a limited range of fields. On the other hand, as for the remaining sections (i.e., A, F, G, and H), the order of their values of $CL_{citing}(n)$ depends on n at which they are compared.

The value of $CL_{citing}(n)$ in section H is very nearly equal to that in section G at the earliest stage where n is small. As n increases, however, the value in H becomes lower than that in G, and then, the disparity between these two sections gradually widens. Furthermore, section H is surpassed by A, after by G, at the early stage; and finally, it is caught up with by F in the value of $CL_{citing}(n)$. These transitions tell us that the growth rate in section H is markedly slow. A slow growth rate of $CL_{citing}(n)$ in synchronic changes means a low possibility that, when an additional citation occurs for a patent under the condition in which the strength of connections between the patent and citing classifications is constant, the classification of the additional citing patent is different from the classifications of previous ones (i.e., a patent receives a citation from another field). That is to say, a patent in section H (electricity) tends to receive citations repeatedly from the same fields as before and to be hardly ever cited by patents belonging to fields different from them even when the cumulative number of citations increases.

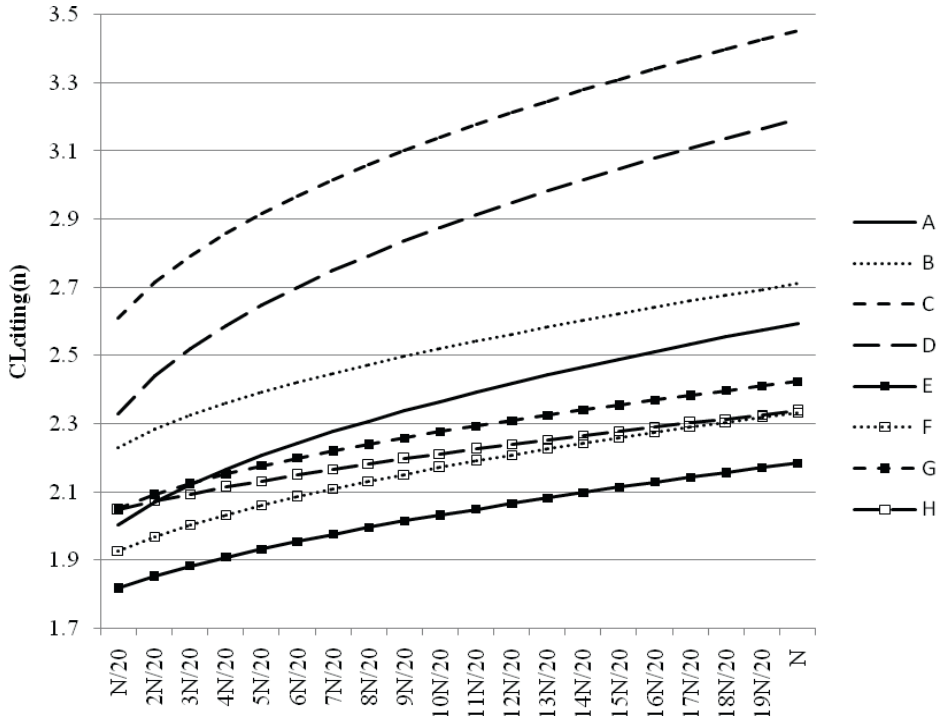

Fig. 3. The developmental profile of $CL_{citing}(n)$ in each section.

There are a few related studies that have examined the assigned classifications in patent citations for each section of IPC (Yoshikane, Suzuki, & Tsuji, 2012; Yoshikane, 2013). These studies reported that a patent giving citations to various fields tends to be cited more frequently, though they did not focus on time-series variation. According to the results shown in Yoshikane, Suzuki, & Tsuji (2012), the correlation coefficient between the number of classifications assigned to backward citations and the number of forward citations is comparatively higher in B (performing operations; transporting), C (chemistry; metallurgy), and D (textiles; paper) than in other sections. On the other hand, as described above, our results show that a patent of these three sections tends to receive citations from a diverse range of fields. Considering these results together, we speculate that, because sections B, C, and D are the fields whose patents receive citations from a wider variety of fields, the tendency that patents giving citations to various fields (i.e., patents based on a variety of technological bases) attract more citations is stronger in these sections than in other sections.

5. Conclusions

This study focused on the classifications assigned to patents, and examined the number of different classifications of patents citing each patent, namely the diversity of citing fields. Specifically, we observed the expected values for growth in the average number of citing fields according to the increase in the data size (i.e., the cumulative number of citations) when regarding the observation period as being in a synchronic state and assuming that the strength of potential connections between each subject patent and citing fields is constant.

The results of comparisons between fields showed that, irrespective of the cumulative number of citations, the transition of the diversity of citing fields is stable with higher values in the fields of “performing operations; transporting,” “chemistry; metallurgy,” and “textiles; paper” (sections B, C, and D), while it is stable with lower values in the field of “fixed constructions” (section E). Moreover, as for the field of electricity (section H), it was

found that the growth rate of the diversity of citing fields is markedly slow—in other words, a patent in this field tends to receive citations repeatedly from the same fields as before and to be hardly ever cited by patents belonging to fields different from them even when the cumulative number of citations increases.

In future research, we would like to examine and clarify changes relating to citation networks of patents and their classifications, not only from the viewpoint of the synchronic change, but also from that of the diachronic change, which is observed through the empirical growth of networks according to the increase in the cumulative number of citations over time.

Acknowledgements

This work was partially supported by Grant-in-Aid for Scientific Research (C) 23500294 (2013) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and I would like to show my gratitude to the support. I also acknowledge Dr. T. Suzuki of the Toyo University and Mr. Y. Suzuki of the University of Tsukuba for their technical advice.

References

Baayen, R. H. (2001). *Word frequency distributions* (Text, Speech, and Language Technology). Dordrecht: Kluwer Academic Publishers.

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika*, *63*(3), 435–447.

Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T., Ehara, T., Echizen-ya, H., & Shimohata, S. (2010). Overview of the patent translation task at the NTCIR-8 workshop. In N. Kando, K. Kishida & M. Sugimoto (Eds.), *Proceedings of the 8th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering, and cross-lingual information access* (pp. 371–376). Tokyo: National Institute of Informatics.

Good, I. J., & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, *43*(1), 45–63.

Haruna, S., Jinji, N., & Zhang, X. (2010). Patent citations, technology diffusion, and international trade: evidence from Asian countries. *Journal of Economics and Finance*, *34*(4), 365–390.

Jang, S. - L., Lo, S., & Chang, W. H. (2009). How do latecomers catch up with forerunners?: analysis of patents and patent citations in the field of flat panel display technologies. *Scientometrics*, *79*(3), 563–591.

Lee, Y. - G. (2009). What affects a patent's value? an analysis of variables that affect technological, direct economic, and indirect economic value: an exploratory conceptual approach. *Scientometrics*, *79*(3), 623–633.

Li, X., Chen, H., Huang, Z., & Roco, M. C. (2007). Patent citation network in nanotechnology (1976–2004). *Journal of Nanoparticle Research*, *9*(3), 337–352.

Li-Ying, J., Wang, Y., Salomo, S., & Vanhaverbeke, W. (2013). Have Chinese firms learned from their prior technology in-licensing?: an analysis based on patent citations. *Scientometrics*, *95*(1), 183–195.

Meyer, M. (2000). What is special about patent citations?: differences between scientific and patent citations. *Scientometrics*, *49*(1), 93–123.

Nanba, H., Fujii, A., Iwayama, M., & Hashimoto, T. (2008). Overview of the patent mining task at the NTCIR-7 workshop. In N. Kando & M. Sugimoto (Eds.), *Proceedings of the 7th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering, and cross-lingual information access* (pp. 325–332). Tokyo: National Institute of Informatics.

Sato, Y., & Iwayama, M. (2006). A study of patent document score based on citation analysis. *Information Processing Society of Japan SIG Technical Report*, *2006*(59), 9–16.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be?: measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323–352.

Wasserman, S., & Faust, K. (1994). *Social network analysis*. New York: Cambridge University Press.

WIPO (World Intellectual Property Organization). (2010). International Patent Classification (IPC), Available: http://www.wipo.int/classifications/ipc/en/

Yoshikane, F. (2013). Multiple regression analysis of a patent's citation frequency and quantitative characteristics: the case of Japanese patents. *Scientometrics*, *96*(1), 365– 379.

Yoshikane, F., & Kageura, K. (2004). Comparative analysis of coauthorship networks of different domains: the growth and change of networks. *Scientometrics*, *60*(3), 433–444.

Yoshikane, F., Suzuki, Y., & Tsuji, K. (2012). Analysis of the relationship between citation frequency of patents and diversity of their backward citations for Japanese patents. *Scientometrics*, *92*(3), 721–733.