Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 73 (2013) 217 – 223

Procedia
Social and Behavioral Sciences

The 2nd International Conference on Integrated Information

Multiple Regression Analysis between Citation Frequency of Patents and their Quantitative Characteristics

Fuyuki Yoshikane[a]*, Yutaka Suzuki[b], Yui Arakawa[b], Atsushi Ikeuchi[a], Keita Tsuji[a]

[a]*University of Tsukuba, Faculty of Library, Information and Media Science, 305-8550, Tsukuba, Japan.*
[b]*University of Tsukuba, Graduate School of Library, Information and Media Studies, 305-8550, Tsukuba, Japan.*

Abstract

Many bibliometric studies have been conducted to examine the factors that influence citation frequency based on multiple regression analysis, targeting academic papers. As for patents, on the other hand, there are few studies in which citation frequency is explained/predicted as the response variable. This study executed a multiple regression analysis that explains citation frequency of patents with multiple feature values derived from the patent data set as explanatory variables, i.e., the numbers of inventors, classifications, pages, figures, tables, claims, priority claims, countries for priority claims, and classifications associated with backward citations (patents cited in the subject patent). The data on 5,253,614 patent applications were analyzed on the basis of the full text of the official gazette for patent applications published in Japan between 1993 and 2007. The results suggested that the influence of diversity of backward citations on citation frequency is large compared to those of the other factors.

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

1. Introduction

The citation frequency of academic papers is thought to express some kind of their importance, and therefore it is often used as a measure in research evaluation. In recent years, the same type of study has been conducted for patents. Of course, similar to the case of citations between papers [1], there are not only positive citations but also negative ones between patents, and their objectives and utility levels vary. However, patent citations are basically perceived as the reutilization of an existing technology [2], and it is thought that a citation denotes the citing patent has found in the cited patent the value to be utilized. In fact, some studies reported relationships between

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

doi:10.1016/j.sbspro.2013.02.044

the importance or economic value of a patent and its citation frequency [3][4]. In other words, to predict the citation frequency of patents or to grasp factors affecting it is meaningful in that it helps us to estimate the importance of patents.

Many studies have been conducted to understand the factors that affect the citation frequency based on the statistical analysis such as multiple regression analysis, targeting academic papers [5][6]. On the other hand, with regard to patent documents, while several studies have tried to distinguish important patents using the citation frequency as one of explanatory variables, there have been few studies in which the citation frequency is explained/predicted as the response variable.

Yoshikane et al. [7] analyzed the relationship between the citation frequency of patents and the diversity of their citations (the number of different classifications associated with patents cited by them). They reported that (i) although the correlation between both was statistically significant, the values of the correlation coefficient were low at approximately 0.1, and that (ii) when patents were grouped by the citation frequency and the diversity of citations was compared among the groups, the diversity of citations in the often cited group was between 1.5 and 4 times as high as that in the less frequently cited group. These results indicate that although the two do not have a simple, linear correlation, there is a possibility that the number of times a patent will be cited (i.e., the number of forward citations) is affected by the diversity of patents which were cited in it (i.e., the diversity of backward citations). However, they paid attention only to the diversity of cited patents as a factor; they did not do an analysis where multiple factors are comprehensively considered. Therefore, we cannot deny the possibility that only an "apparent correlation," where multiple factors are confounded, is shown.

With this as the background, we have executed a multiple regression analysis that explains the citation frequency with multiple feature values that can be derived from the patent data set as explanatory variables in this study. In addition to the number of classifications assigned to backward citations, we included eight variables; the numbers of inventors, pages, claims, and so on.

2. Data

The NTCIR test collections compiled by the National Institute of Informatics (NII), Japan were our information sources, 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 [8] and the 1,757,361 documents published in the five years between 2003 and 2007 from NTCIR-8 Patent Translation Test Collection [9]. Approximately 350,000 documents were published in each of these years. The targets of the analysis are the 341,388 patent applications published in 1998. We investigated the classifications of patents cited by them and the number of times each of them is cited among the 10 years following their publication, that is, from 1998 to 2007. As for the classifications assigned to cited patents, the investigation was based on the "patent gazette (publication of unexamined patent applications)" published in the period where the data were available, that is, from 1993 and later. Therefore, if a patent published in 1992 or earlier was cited, its classifications are not able to be identified. But even though there is this limitation, we consider it reasonable to assume that it does not have a serious effect on the results, which will reveal general tendencies of citation among patents.

There are cases in which the descriptions of cited patents are inserted in the main text of the patent document, rather than as independent items, and moreover the format of those descriptions is not standardized. Thus, it is difficult to completely and precisely extract the information of cited patents, particularly for the patents published before 2002 when the information disclosure system for prior art documents was introduced in Japan [10]. Furthermore, there are numerous instances of typographical errors that are assumed to have occurred with digitization of patent documents, such as mistakes in the software conversion of Kana (Japanese phonetic alphabets) to Kanji (ideographic characters) and those in optical character recognition conversion [2]. Similar problems exist also in extracting or calculating some of the feature values of the patent application used in the multiple regression analysis as explanatory variables, which will be indicated in the next section. This study has dug and listed the

variants of description patterns in patent documents through data observation, and covered those variations in obtaining values of each variable.

3. Methods

We executed a multiple regression analysis based on the linear regression model using the following variables.

3.1. Response variable

The response variable is the number of times the subject patent is cited by others, namely the citation frequency, during the ten years after its publication (F_{cited}). Figure 1 shows the distribution of the citation age, which means how many years later a patent is cited after its publication (the difference between cited and citing patents not for one's publication year and the other's application year, but for both publication years). The distribution is expressed per "section," which is the top layer in the International Patent Classification (IPC). Section A is "human necessities," B is "performing operations; transporting," C is "chemistry; metallurgy," D is "textiles; paper," E is "fixed constructions," F is "mechanical engineering, etc.," G is "physics," and H is "electricity" [11]. In general, the distribution of the citation age has similar tendencies for all the sections. Regarding section D, however, the proportion of citations in the period just after publication (during a few years) is small compared to the other sections. For all the sections, there are still many citations even after nine years. So, in order to get an overall picture of the citation behavior for patents, it would be necessary to observe data over a longer time period. However, after being peaked at around six years later, the frequency of citations begins to decrease. We consider that covering ten years in the observation would provide at least an understanding in broad outline.

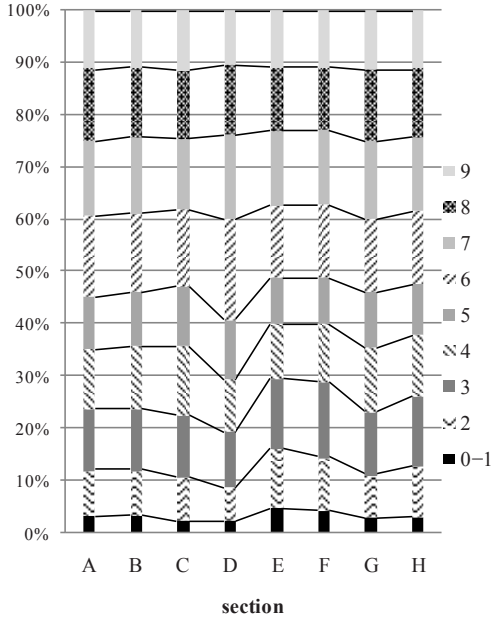

Fig. 1. Citation age of patents for each section

3.2. Explanatory variable

As for explanatory variables, we adopt the numbers of inventors (*IV*), associated classifications (*VC*), pages (*PG*), figures (*FG*), tables (*TB*), claims (*CL*), priority claims (*PC*), countries for priority claims (PC_c), and classifications associated with the patents that the subject patent is citing (VC_{citing}).

While the number of inventors for a patent application corresponds to the number of authors for an academic paper, the numbers of pages, figures, and tables represent the quantities of descriptions in a document. These indices have been dealt with in the correlation or regression analyses of the citation frequency targeting academic papers [12][13][14]. On the other hand, the number of claims, number of priority claims, and number of countries for priority claims are quantities specific to patent applications, which are related to rights. The number of classifications of the subject patent and the number of classifications of patents cited by it (its backward citations) are quantities that reflect the diversity of the invention's contents, and the relationship with the citation frequency of the subject patent, i.e., the number of times it is cited by others (its forward citations) has been pointed out [7].

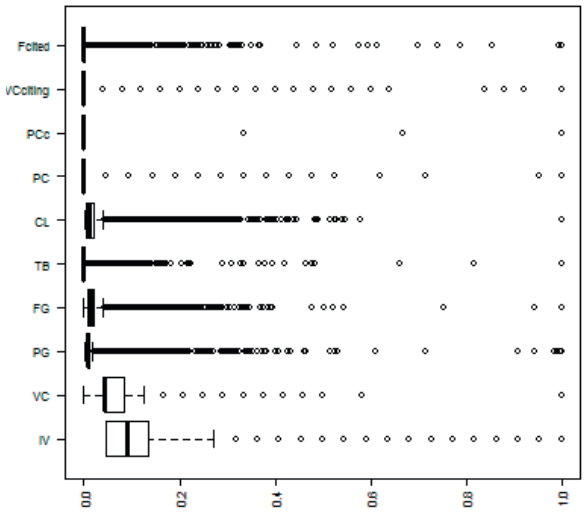

Fig 2. Distribution of values for each index.

Since all of these indices have highly skewed distributions, we have transformed each index, x, into natural logarithmic values, $ln(x+1)$, before applying them as response/explanatory variables to the multiple regression analysis. We added one to the values for avoiding zero in the logarithmic transformation. Figure 2 is a box plot that shows for each of the indices the distribution of values, which are normalized to [0, 1] by dividing its maximum value. The right side of each box shows the 75th percentile value. Regarding other than the number of inventors (*IV*) and the number of associated classifications (*VC*), those values are concentrated in the area around each minimum value, and we can confirm that these two (*IV* and *VC*) also have very skewed distributions, where the 75th percentile value is no more than about 1/10th of the maximum value. Setting the variable inclusion criteria at the statistically significant probability value (p-value) $p_{in} < 0.05$, and the variable exclusion criteria at $p_{out} > 0.05$, we introduced explanatory variables selected through the stepwise method.

4. Results

Table 1 shows the results of the multiple regression analysis executed for each section of A−H as well as those for all of the patents published in 1998. The following values are shown in the table: the number of subject patents n, the coefficient of determination adjusted for degrees of freedom R'^2, and the standardized partial regression coefficient for each explanatory variable. Since it is common to assign multiple classifications to one patent, the sum of n under the classifications from A to H is greater than the total n of patents (the bottom row of Table 1). Variables not selected by the stepwise method are denoted as "*N.S.*"

Not only regarding the whole data of patents being studied but also regarding any section, all of the regression were statistically significant ($p < 0.001$). However, the values of the coefficient of determination R'^2 were low, that is to say, the regression model was not well fitted. In the field of D (textiles; paper), R'^2 was higher than in the other fields, but its value was not more than 0.10. As with the coefficient of determination, the absolute values for the standardized partial regression coefficient were low in general. The number of pages of the subject patent (PG) and the number of classifications of the patents cited in the subject patent (VC_{citing}) had high values of the regression coefficient, compared with the other explanatory variables. In particular, for the number of classifications of the cited patents, the values were kept relatively high throughout most sections.

The number of classifications of the subject patent (VC), as well as that of the cited patents (VC_{citing}), had significant and positive values of the regression coefficient in all the sections except A (human necessities) ($p < 0.001$). It is observed that the regression coefficient tends to be higher for VC_{citing} than for VC. This result, which shows the diversity in the classification distribution for the patents cited in the subject patent exhibits stronger relationships with the citation frequency (F_{cited}) than that for the subject patent itself does, falls in with the results reported by Yoshikane et al. [7]. The number of inventors (IV), number of tables (TB), and number of claims (CL) had positive values of the regression coefficient at the same level as the number of associated classifications (VC). Although the number of figures (FG) had negative values of the regression coefficient, it was not selected in some sections. As for the number of priority claims (PC) and number of countries for priority claims (PC_c), in half of the sections either or both were not selected by the stepwise method. It would be caused by that the two variables were highly correlated with each other.

Table 1. Results of the multiple regression analysis

	n	R'^2		
			IV	***VC***
A	39474	0.035*	0.046*	*N.S.*
B	92483	0.050*	0.045*	0.050*
C	46881	0.030*	0.046*	0.029*
D	6255	0.073*	0.070*	0.070*
E	22564	0.015*	0.036*	0.041*
F	40437	0.029*	0.049*	0.050*
G	103919	0.043*	0.049*	0.047*
H	96679	0.037*	0.056*	0.055*
Whole	341388	0.038*	0.049*	0.036*

	Standardized partial regression coefficient		
	PG	***FG***	***TB***
A	0.016	−0.038*	0.097*
B	0.098*	−0.058*	0.057*
C	0.048*	−0.023*	0.042*
D	0.067*	−0.121*	0.072*
E	0.030*	*N.S.*	0.038*
F	0.092*	*N.S.*	0.031*
G	0.086*	−0.044*	0.046*
H	0.071*	−0.023*	0.040*
Whole	0.073*	−0.039*	0.057*

	CL	***PC***	PC_c	VC_{citing}
A	0.060*	−0.063*	*N.S.*	0.071*
B	0.047*	0.157*	−0.164*	0.077*
C	0.045*	0.069*	−0.088*	0.098*
D	0.031	*N.S.*	−0.038	0.068*
E	0.053*	*N.S.*	*N.S.*	0.031*
F	0.028*	*N.S.*	−0.028*	0.063*
G	0.078*	0.107*	−0.132*	0.086*
H	0.076*	0.104*	−0.129*	0.071*
Whole	0.063*	0.067*	−0.095*	0.075*

5. Conclusions

In this study, we executed a multiple regression analysis using nine factors in patent applications as explanatory variables, i.e., the numbers of inventors, associated classifications, pages, figures, tables, claims, priority claims, countries for priority claims in a patent, and the number of classifications associated with the patents cited in it, for the purpose of examining the influence of these factors on the citation frequency (i.e., the number of forward citations). Comparing the standardized partial regression coefficient of each variable, the number of classifications associated with cited patents (i.e., the diversity of backward citations) had higher values, regardless of technological fields. That is to say, even taking into account confounding with other factors, we could obtain the results which suggest an influence of the diversity of backward citations in a patent on the number of its forward citations.

We concluded that the values of the standardized partial regression coefficient for the number of classifications associated with cited patents were relatively high: nevertheless, this is not anything more than a result based on the comparison to other variables. Although they were statistically significant, the values themselves did not necessarily reach a meaningful level. This is probably caused by the fact that because in most patent applications the value of the citation frequency, which was used as the response variable, was zero as described in Section 3, it is difficult to apply linear regression to this type of data. Setting threshold values for the citation frequency, we will clarify the contribution of each factor in discriminating patents whose citation is above the threshold, on the basis of non-linear or generalized linear models such as logistic regression, in future studies.

References

[1] Bornmann, L., & Daniel, H. -D. (2008). What do citation counts measure?: a review of studies on citing behavior. Journal of Documentation, 64(1), 45-80.
[2] Inuzuka, A. (2011). Factors facilitating technology reuse: an estimation from patent citation data. Okayama Economic Review, 43(3), 15-28.
[3] Narin, F. (1995). Patents as indicators for the evaluation of industrial research output. Scientometrics, 34(3), 489-496.
[4] Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions, The Review of Economics and Statistics, 81(3), 511-515.
[5] Peters, H. P. F., & van Raan, A. F. J. (1994). On determinants of citation scores: a case study in chemical engineering. Journal of the American Society for Information Science, 45(1), 39-49.
[6] Bornmann, L., & Daniel, H. -D. (2007). Multiple publication on a single research study: does it pay? the influence of number of research articles on total citation counts in biomedicine. Journal of the American Society for Information Science and Technology, 58(8), 1100-1107.
[7] 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.
[8] Nanba, H., Fujii, A., Iwayama, M., & Hashimoto, T. (2008). Overview of the patent mining task at the NTCIR-7 workshop. Proceedings of NTCIR-7 Workshop Meeting, Tokyo, Japan, 325-332.
[9] Fujii, A., Utiyama, M., Yamamoto, T., Utsuro, T., Ehara, H., Echizen-ya, H., & Shimohata, S. (2010). Overview of the patent translation task at the NTCIR-8 workshop. Proceedings of NTCIR-8 Workshop Meeting, Tokyo, Japan, 371-376.
[10] 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.
[11] WIPO (World Intellectual Property Organization). (2010). International Patent Classification (IPC), Available: http://www.wipo.int/classifications/ipc/en/
[12] Snizek, W. E., Oehler, K., & Mullins, N. C. (1991). Textual and nontextual characteristics of scientific papers: neglected science indicators. Scientometrics, 20(1), 25-35.
[13] Glänzel, W. (2002). Co-authorship patterns and trends in the sciences (1980-1998): a bibliometric study with implications for database indexing and search strategies,.Library Trends. 50(3), 461-473.
[14] Kostoff, R. N. (2007). The difference between highly and poorly cited medical articles in the journal Lancet. Scientometrics, 72(3), 513-520.